Implementation of a Super-Resolution Ambisonics-to-Binaural Rendering Plug-In

Project Thesis

Supervisor: Dipl. Ing. Christian Schörkhuber

Peter Maximilian Giller

Graz, January 24, 2019

institut für elektronische musik und akustik

Abstract

Binaural reproduction of Ambisonic sound scenes is an active research area. Potential applications are to be found in gaming and virtual reality, simulation and auralization, or music production. Fundamentally different approaches for binaural rendering exist. The well-known least-squares decoder suffers from poor resolution and externalization, and, most notably, a severe roll-off towards higher frequencies if the order of the input signal is low. The latter is due to the significant amount of energy contained in HRTFs in higher-order modes. While various methods exist to remedy timbral artifacts or to enhance the spatial resolution, the benefit of signal-independent rendering methods appears to be limited for lower-order input signals. A signal-dependent binaural renderer in form of an open-source audio plug-in is presented in this thesis. The plug-in implements a recently published method which is a parametric extension of the least-squares decoder, whereby direct sound impinging from the most prominent direction of arrival is reproduced exactly, and diffuse sound is reproduced unaltered in terms of the stochastic parameters.

Contents

1 Introduction

Immersive technologies are on the rise. The term *immersion* describes the sensation of losing oneself in an artificially created environment while the level of awareness of the underlying reality diminishes. While the visual component tends to attract the main focus of research as well as the public attention, virtual three-dimensional sound scenes are an integral part of the experience, and can also stand for themselves. Potential applications can be found in gaming and virtual reality, simulation and auralization, 3D music recording and production, or teleconferencing.

Because of the expense and space requirements of loudspeaker arrays, it is desirable to present 3D audio over headphones, or *binaural*, while the perceived position and orientation of the listener in the sound field are adapted according to head or body movements. Binaural audio requires to emulate the very physical cues that evoke the spatial impression, and make use of psychoacoustic effects that occur when sound waves impinge on the ears.

By which criteria can the authenticity of an acoustic virtual reality experience be evaluated? Depending on the application, a prerequisite is that point-like sound sources must be presented in a way that allows precise localization. The listener should furthermore be able to judge the width and distance of sound sources, and get a natural impression of the virtual space, i.e., early reflections and diffuse reverberation. Interactive response to head movements within the virtual environment increases the immersive effect. Finally, any quality impairments or artifacts, e.g., sound colorations, need to be avoided, as they potentially draw the observer's attention to the presence of the employed technology.

Ambisonics is a widely used and flexible scene-based format for 3D audio. Ambisonic signals can be directly recorded or obtained from an object-based audio representation. Binaural decoding is particularly interesting for Ambisonic input signals because it easily allows for coordinate rotations; Ambisonics is therefore well suited for virtual reality applications. The orthogonality and completeness (assuming sufficiently high order) of the Ambisonic sound field description in the angular domain makes it convenient for the application of various transformations, spatial filtering or source extraction. The problem of Ambisonics-to-binaural decoding is subject to intensive research.

Due to restrictions of feasible microphone arrays and limited computational power, it is difficult to obtain high-quality Ambisonic recordings of natural sound fields with high order. Over the past few years, different methods to increase the spatial resolution of low-order Ambisonics have been proposed. Some include binaural decoding; if not, they can be employed as an intermediate step before binaural decoding with established methods.

Unfortunately, many rendering approaches suffer from severe artifacts, such as the roll-off towards high frequencies for low input orders of the least-squares decoder. Despite recent progress in the field of Ambisonic-to-binaural rendering, this problem remains especially for data-independent methods. Data-dependent rendering can further improve the results; not only by reducing artifacts, but also by trying to increase the directional

sharpness beyond the resolution of the input signal.

This thesis focuses on a data-dependent super-resolution binaural rendering method that has recently been proposed. This method tries to increase the directional sharpness while preserving the perceptual qualities of the original sound field, such as the impression of diffuseness and spaciousness. Similarly to Directional Audio Coding (DirAC), the method is based on frequency-bin-wise estimation of the directional of arrival; however, the diffuse part is reproduced correctly without the need to estimate the signal-to-diffuse ratio.

The main contribution of this work is the efficient implementation of an Ambisonics-to-binaural renderer that applies the new method for real-time rendering, and the pre-computation of the filter weights for arbitrary HRTF sets. The renderer is implemented in C++ as an audio plug-in that can be used in any common digital audio workstation. However, due to object-oriented design, the algorithm source code can easily be integrated into other projects.

Section 2 provides a theoretical basis, starting with Ambisonics and the spherical harmonic transform. After a subsequent introduction to localization and binaural audio, a short review of different signal-independent binaural rendering methods will be provided. Three different approaches to enhance the resolution of Ambisonics signals will be presented in Section 2.4.

Section 3 derives the new signal-dependent method after expounding the signal model and the underlying assumptions.

The pre-computation of the filter weights and implementation-specific details will be discussed in Section 4.

This thesis closes with a summary in Section 5.

Notice

Parts of this thesis have been published subsequently, including a listening experiment comparing different binaural rendering strategies. Please refer to [1] for the most recent version of this project.

2 Theoretical Background

2.1 Ambisonics

Ambisonics is a spatial audio format that relies on the spatial decomposition of a sound field. Rather than describing the loudspeaker feeds for a fixed playback configuration (channel-based), it holds the components which together add up to the original sound field at a certain point in space (scene-based). Ambisonics is not bound to a fixed playback setup like, e.g., Dolby 5.1 is. The requirements regarding the playback setup rather concern the total number of loudspeakers and their distribution. Furthermore, Ambisonics allows for various transformations, such as rotation, warping of the coordinate space, or directional loudness modifications [2]. Ambisonic signals can be obtained by recording a real sound field with a microphone array, or by superposition of direct encoded point-like sources.

The concept of Ambisonics is strongly related to the spherical harmonic decomposition or spherical Fourier transform which will be also explained later on.

2.1.1 Ambisonic Signal Representation

Ambisonics is based on the spatial decomposition of two- or three-dimensional sound fields at the surface of a sphere. The extent of decomposition, indicated by the Ambisonics order, determines the angular resolution. The decomposition into components up to a certain order N includes all lower-order components $0 \leq n \leq N$.

First Order Ambisonics (FOA), also known as *B-Format*, was first described by Michael A. Gerzon in the 1970s [3]. It consists of four channels which can be thought of as the output of a set of microphones located at the coordinate origin that have either monopole or dipole characteristics. The zeroth order component, usually referred to as the W signal, corresponds to the output signal of an omnidirectional microphone, whereas the first order components X, Y, and Z, each correspond to a figure-of-eight characteristic aligned with the respective coordinate direction. The described configuration can already be used to make Ambisonic recordings. However, it is more common to obtain the four B-Format signals by a linear combination of the outputs of an array consisting of four cardioid microphones at the corners of a regular tetrahedron. In this case, the distance of all four capsules of a tetrahedral array to the center is the same, and the spherical sampling is optimal (cf. Section 2.1.2).

The microphone characteristics of the four FOA channels form an orthonormal basis on the sphere. Every incoming sound wave is projected onto this basis, i.e., decomposed into a linear combination of these functions. Note that the purely angular formulation assumes the incident waves to be plane, and therefore discards any distance information.

The superposition of the omnidirectional component and each of the dipoles results in a cardioid characteristic. Since achievable polar patterns are restricted to cardioid-like characteristics, it is intuitively clear that spatial resolution of FOA is limited.

Encoding of a point source to a direction Ω_0 in FOA can be achieved if the point source signal is weighted by the respective basis function evaluated at Ω_0 for each Ambisonics channel.

The concept of Ambisonics can be extended to arbitrary orders if the sound field is further decomposed. Higher-order Ambisonics is based on the projection onto additional basis functions with increased directional selectivity. A suitable set of basis functions is formed by the so-called *spherical harmonics* (SHs).

The real-valued spherical harmonics of order n and degree m are given by

$$Y_n^m(\Omega) = \begin{cases} C_n^m \cos(m\varphi) P_n^m(\cos\vartheta) & \text{if } m \geq 0 \\ C_n^{|m|} \sin(|m|\varphi) P_n^{|m|}(\cos\vartheta) & \text{if } m < 0 \end{cases}, \; n \in \mathbb{N}_0,\, m \in \mathbb{Z},\, |m| \leq n, \tag{1}$$

where C_n^m is a normalization term, P_n^m is the associated Legendre polynomial, and $\Omega = (\vartheta, \varphi)$ is the set of angular coordinates on the unit sphere. $\vartheta \in [0, \pi)$ is the elevation angle starting at the negative z axis, and $\varphi \in [0, 2\pi)$ is the azimuth angle which is zero in positive x direction and increases towards the y axis.

The spherical harmonics are orthogonal, but the normalization term C_n^m can be chosen with respect to the convention used in order to ensure they are also orthonormal, i.e., the inner product

$$\int_{\Omega \in \mathcal{S}^2} Y_n^m(\Omega) Y_{n'}^{m'}(\Omega) \mathrm{d}\Omega = \delta_{nn'} \delta_{mm'} \tag{2}$$

of two spherical harmonics on the surface of the unit sphere, denoted by $\mathcal{S}^2$, where $\int_{\Omega \in \mathcal{S}^2} \mathrm{d}\Omega = \int_{-\pi}^{\pi} \int_0^{\pi} \mathrm{d}\varphi \cos\vartheta \mathrm{d}\vartheta$, is one if they are of the same order and degree, and zero otherwise. δ_{ij} is the Kronecker delta function,

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \tag{3}$$

We define an Ambisonics signal of order N as a vector of Ambisonic components

$$\boldsymbol{z}(t) = \begin{bmatrix} z_0^0(t) & z_1^{-1}(t) & z_1^0(t) & z_1^1(t) & \cdots & z_N^{-N}(t) & \cdots & z_N^N(t) \end{bmatrix}^\mathsf{T} \tag{4}$$

that each correspond to to a signal recorded by a virtual microphone with directivity pattern corresponding to the spherical harmonic of same order and degree.

2.1.2 Spherical Harmonic Transform

Any square-integrable function on the sphere $f(\Omega)$ can be decomposed into a weighted sum of spherical harmonics

$$f(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_n^m(\Omega) \tag{5}$$

with coefficients f_{nm}. The coefficients f_{nm} can be computed by the inner product of $f(\Omega)$ and the spherical harmonics Y_n^m:

$$f_{nm} = \int_{\Omega \in \mathcal{S}^2} f(\Omega) Y_n^m(\Omega) \mathrm{d}\Omega. \tag{6}$$

Due to its equivalence to a Fourier series expansion at the unit sphere, the spherical harmonic decomposition is also known as spherical Fourier transform.

In practice, the order can not be chosen close to infinity. Not only are audio computations at very high orders highly expensive; moreover, in the case of spherical microphone array processing, the maximum achievable order is very limited. Most of the Ambisonics recordings are restricted to orders below 5 due to the difficulty of placing a large number of high-quality microphones within a sufficiently small array. First-order Ambisonics is by far the most widespread case. Hence, the series expansion from (5) is usually truncated at a certain order N:

$$f(\Omega) \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} f_{nm} Y_n^m(\Omega) \tag{7}$$

In analogy to the Nyquist-Shannon sampling theorem, perfect reconstruction of f is now only possible if f is band-limited in the angular domain, i.e., the maximum order of f does not exceed N. Natural sound fields are, in general, not order-limited. Therefore, sampling may introduce aliasing of higher-order components into lower orders, depending on the spherical wavelength and the aperture of the microphone array.

Given an order-limited sound field, the maximum possible order of the SH decomposition depends on the number of samples available, as well as their distribution on the sphere. Sampling approximates analysis integral (6) by a discrete sum,

$$f_{nm} = \sum_{j=1}^{J} g_j f(\Omega_j) Y_n^m(\Omega_j), \tag{8}$$

where $\{f(\Omega_j)\}$ is a finite set of samples on the sphere at positions Ω_j. The quadrature weights g_j are introduced to support the approximation. Exact sampling (perfect reconstruction of $f(\Omega)$) requires to maintain orthonormality. Substituting (5) in (8), the modified orthonormality criterion becomes (cf. [4]):

$$\sum_{j=1}^{J} g_j Y_{n'}^{m'}(\Omega_j) Y_n^m(\Omega_j) = \delta_{nn'}\delta_{mm'} \tag{9}$$

Among the different sampling schemes – to name but a few, e.g., equi-angle, spiral, or Gaussian sampling – those are of particular interest which lead to an approximation that is proportional to the analysis integral. *Spherical t-designs* provide a set of coordinates $\mathcal{M}$ on the unit sphere that guarantees orthogonality, i.e.,

$$\sum_{\Omega_j \in \mathcal{M}} Y_{n'}^{m'}(\Omega_j) Y_n^m(\Omega_j) \propto \delta_{nn'}\delta_{mm'}, \tag{10}$$

and allows to replace the integral of a t-th order polynomial $\mathcal{P}_t$ on the sphere surface $\mathcal{S}^2$ by a discrete sum with equal weights [5]

$$\int_{\Omega \in \mathcal{S}^2} \mathcal{P}_t(\Omega) \mathrm{d}\Omega = \frac{4\pi}{J} \sum_{\Omega_j \in \mathcal{M}} \mathcal{P}_t(\Omega_j). \tag{11}$$

In order to formulate the spherical harmonic transform in matrix notation, we express $\{f(\Omega_j)\}$ by a vector $\boldsymbol{f} = [f(\Omega_1) \ \cdots \ f(\Omega_J)]^\mathsf{T}$ and denote the discrete spherical Fourier transform of $\boldsymbol{f}$ by

$$\mathring{\boldsymbol{f}}_N = \mathcal{SHT}\{\boldsymbol{f}\}, \tag{12}$$

where $\mathring{\boldsymbol{f}}_N = [f_{00} \ f_{1\,-1} \ f_{10} \ f_{11} \ \cdots \ f_{NN}]^\mathsf{T}$. The forward spherical harmonic transform in matrix notation for the general case is given by

$$\mathring{\boldsymbol{f}}_N = \boldsymbol{Y}_N \, \boldsymbol{G} \, \boldsymbol{f}, \tag{13}$$

which is equivalent to the analysis sum (8), where the matrix of order-N spherical harmonics is $\boldsymbol{Y}_N = [\boldsymbol{y}(\Omega_1) \ \cdots \ \boldsymbol{y}(\Omega_J)]$ for directions $\{\Omega_j\}$ with $\boldsymbol{y}_N(\Omega) = [Y_0^0(\Omega) \ Y_1^{-1}(\Omega) \ \cdots \ Y_N^N(\Omega)]^\mathsf{T}$, and $\boldsymbol{G} = \mathrm{diag}\{g_i\}$. The inverse spherical harmonic transform (ISHT) of $\mathring{\boldsymbol{f}}$ is given by

$$\boldsymbol{f} = \mathcal{ISHT}\{\mathring{\boldsymbol{f}}_N\} = \boldsymbol{Y}_N^\mathsf{T} \, \mathring{\boldsymbol{f}}_N. \tag{14}$$

Again, (13) depends on the choice of $\boldsymbol{G}$ and the sampling locations. The least squares solution is given by

$$\mathring{\boldsymbol{f}}_N = \boldsymbol{Y}_N^\dagger \, \boldsymbol{f} \ , \tag{15}$$

where $\boldsymbol{Y}_N^\dagger$ is the Moore-Penrose pseudoinverse of $\boldsymbol{Y}_N^\mathsf{T}$. The orthogonality criterion translated into matrix notation is given by

$$\boldsymbol{Y}_N^{\mathsf{H}} \boldsymbol{G}^{\mathsf{H}} \boldsymbol{G} \boldsymbol{Y}_N \propto \boldsymbol{I},\ \Omega_j \in \mathcal{M}. \tag{16}$$

If the set of coordinates constitutes a t-design, and if $\boldsymbol{Y}_N = [\boldsymbol{y}_N(\Omega_j)]_{\Omega_j \in \mathcal{M}}$, and $\boldsymbol{f} = [f(\Omega_j)]_{\Omega_j \in \mathcal{M}}$, the orthogonality criterion holds for unit weights

$$\boldsymbol{Y}_N^{\mathsf{H}} \boldsymbol{Y}_N \propto \boldsymbol{I},\ \Omega_j \in \mathcal{M}, \tag{17}$$

and the solution

$$\mathring{\boldsymbol{f}}_N = \mathcal{SHT}\{\boldsymbol{f}\} = \boldsymbol{Y}_N\, \boldsymbol{f} \equiv \boldsymbol{Y}_N^{\dagger}\, \boldsymbol{f}\ ,\ \Omega_j \in \mathcal{M} \tag{18}$$

is optimal in the least-squares sense. The Moore-Penrose pseudoinverse can be determined via singular value decomposition (SVD):

$$\boldsymbol{Y}_N^{\mathsf{T}} = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^{\mathsf{T}} \tag{19}$$

In the above equation, $\boldsymbol{U}$ and $\boldsymbol{V}$ contain the left- and right-eigenvectors, respectively, and $\boldsymbol{S}$ is a diagonal matrix containing the singular values of $\boldsymbol{Y}_N^{\mathsf{T}}$. The pseudoinverse of $\boldsymbol{Y}_N^{\mathsf{T}}$ is then given by

$$\boldsymbol{Y}_N^{\dagger} = \boldsymbol{U} \boldsymbol{S}^{-1} \boldsymbol{V}^{\mathsf{T}}. \tag{20}$$

In many cases, the inverse of $\boldsymbol{S}$ is ill-conditioned due to singular values which are close to zero. A numerically stable inverse can be obtained using Tichonov regularization

$$\boldsymbol{Y}_N^{\dagger} = \boldsymbol{U} (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1} \boldsymbol{V}^{\mathsf{T}}, \tag{21}$$

or by discarding all singular values that fall below a certain threshold (truncated SVD). The result is the regularized least squares solution

$$\boldsymbol{Y}_N^{\dagger} = \boldsymbol{U} \begin{bmatrix} \tilde{\boldsymbol{S}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^{\mathsf{T}}, \tag{22}$$

where $\tilde{\boldsymbol{S}}$ contains the singular values that lie above the threshold.

2.2 Binaural Audio

The need for relatively large and expensive loudspeaker arrays for playback is a downside of Ambisonics (and other high-resolution surround techniques). Thus, it is desirable to find a way to decode Ambisonics for binaural presentation, rather than for loudspeakers, without altering the impression of sound, space, and diffuseness. But beyond being only a workaround for listening to Ambisonics recordings without having access to adequate loudspeaker arrays, combining binaural techniques with Ambisonics enables us to create rotatable sound scenes and realistic virtual spaces via earphones. The following section outlines the mechanisms of localization and binaural audio. In the following, different binaural rendering approaches are presented.

2.2.1 Localization

Localization refers to the ability to judge the positions of sound sources in geometrical space relative to the listener's head position in a natural or artificial sound field. Given the case of a natural sound field, the sound sources are perceived to be located outside the listener's own head (*externalized*). *Inter-aural time differences* (ITD) and *inter-aural level differences* (ILD) are considered to be the predominant cues for localization. ITDs describe the difference in arrival time of incident sound waves that results from the different path lengths from the source to each of the two ears. Because of the phase ambiguity at higher frequencies, ITDs are expected to provide localization cues for low frequencies, while ILDs rather contribute to localization at high frequencies due to the diffraction around the head at lower frequencies.

Purely virtual phantom sources can be created by a combination of two (or more) loudspeaker signals if the level and delay differences of each signal relative to the other signals are manipulated. The relationship between panning configurations and the perceived location is complex and shall not be treated in this thesis. However, if the loudspeakers are positioned close to or inside the auditory canal or, more specifically, if earphones are used, phantom sources are perceived to be located on the inter-aural (lateral) axis inside the head. This phenomenon is the most important problem binaural techniques need to overcome.

In order to manipulate audio signals in such a way that they are externalized, we need to identify and apply the very cues our hearing evaluates for localization of physical sound sources. While the Duplex Theory [6] finds that ILDs and ITDs are sufficient for localization, it does not explain the localization of sources that have the same ILD and ITD. These points are located at the mantle of a cone symmetric around inter-ear axis with its apex at the center of the head (the so-called *Cones of Confusion*). The fact that we are able to discriminate between these positions is, besides possible visual cues, due to reflections at pinnae, and torso (which are not axisymmetric), that contribute to the direction-dependent frequency response from the source to each of the ears. This leads to frequency-dependent ITDs and a more complex frequency dependency of ILDs compared to the Duplex Theory.

2.2.2 Head-Related Transfer Function

Binaural techniques make use of the assumption that sound waves entering the ear canal already contain all relevant directional cues. Consequently, a sound scene which is first recorded and then played back inside or close to the auditory canal of the very same human subject should be perceived to be identical to the original sound scene (neglecting the impact of microphone and loudspeaker on the recorded signal). A naive approach to obtain a binaural recording could be to provide the target listener with in-ear microphones, which is impractical for obvious reasons. Using *artificial* or *dummy heads* instead turns out to work reasonably well for many listeners, despite the inter-individual differences in geometry of pinnae, head, and torso.

In order to encode virtual sources for binaural presentations, we apply frequency-dependent ITDs and ILDs in the form of a filtering operation with the *head-related transfer functions* (*HRTF*s). A HRTF characterizes the transfer path from a source position to each of the two ears by means of frequency-dependent gains and phase shifts. HRTFs can be measured for individual subjects and also for artificial heads. A virtual source can be positioned if a signal is convolved with the *head-related impulse response* (HRIR) that was measured from the desired direction. If the HRTF for a certain direction is not available, adjacent HRTFs may be interpolated.

2.3 Signal-independent Binaural Rendering Methods

Now that we learned how to encode a monophonic signal for binaural audio, how can we apply this knowledge to more complex signals like Ambisonics recordings that probably contain a large number of direct sources, reflections, and reverberation from arbitrary directions? Obviously, direct encoding of *all* partial sources is not feasible since it would involve the extraction of an unimaginable number of partial sources and very likely produce severe artifacts.

In this section, the two most widely used signal-independent binaural rendering approaches will be presented: the virtual loudspeaker decoder and the least squares decoder. Subsequently, different recently published strategies are presented which strive to reduce the artifacts typically introduced the latter method.

2.3.1 Binaural Rendering Using Virtual Loudspeakers

Fig. 1 shows the block diagram of a widely used binaural decoder that is based on a very straightforward approach. Instead of decoding an N-th order Ambisonics input signal $\boldsymbol{z} = [z_0^0 \ \dots \ z_N^N]^\mathsf{T}$ for a physically present loudspeaker array, it is decoded for a set of L virtual loudspeakers. Each virtual loudspeaker feed s_l is convolved with the left and right HRIR evaluated for the direction of each loudspeaker, or multiplied by the HRTF coefficients in frequency-domain

$$\boldsymbol{h}(\Omega_l, \omega) = [H_\mathsf{L}(\Omega_l, \omega) \ H_\mathsf{R}(\Omega_l, \omega)]^\mathsf{T} \ \ l = 1, \dots, L.$$

Thus, the resulting output signals for the left and right ear

$$\hat{\boldsymbol{x}}(\omega) = [\hat{x}_\mathsf{L}(\omega)\ \hat{x}_\mathsf{R}(\omega)]^\mathsf{T} = \sum_{l=1}^{L} s_l(\omega)\, \boldsymbol{h}(\Omega_l, \omega)$$

are the sum of all virtual loudspeaker signals filtered with the left or right HRTF, respectively.

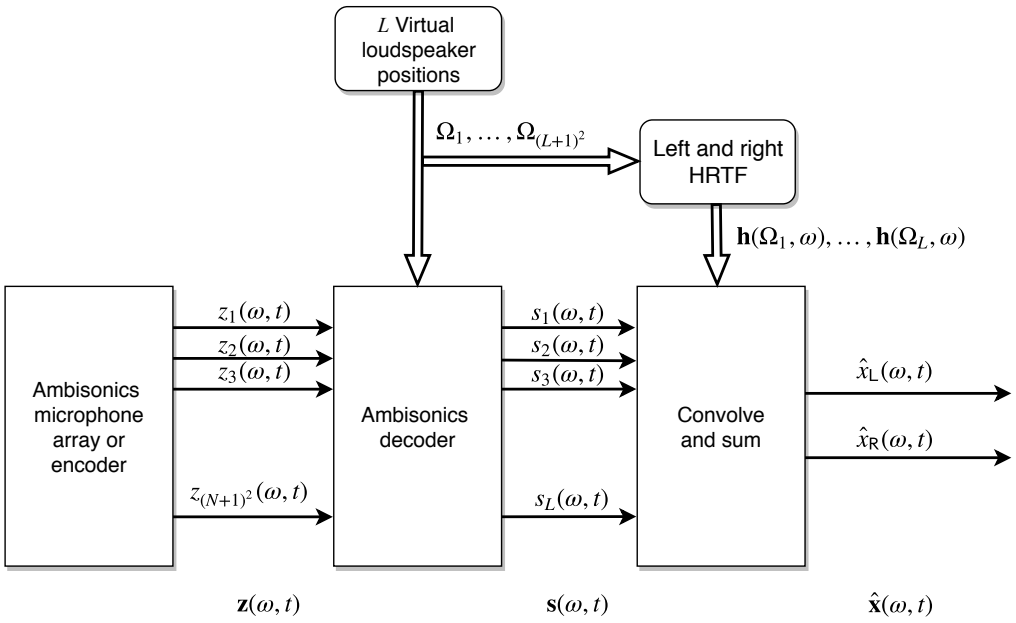


Figure 1: Binaural rendering using virtual loudspeakers.

2.3.2 The Least Squares Decoder

The Ambisonics decoder and the virtual loudspeaker setup have a major impact on the results of the binaural rendering method presented in the previous section. A bad choice can introduce colorations or reduce the spatial resolution. Since only linear operations are involved, the method could be conflated into a single multiple-input multiple-output filtering operation

$$\hat{\boldsymbol{x}}(\omega) = \begin{bmatrix} \hat{x}_\mathsf{L} \\ \hat{x}_\mathsf{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_\mathsf{L}^\mathsf{H} \boldsymbol{z} \\ \boldsymbol{w}_\mathsf{R}^\mathsf{H} \boldsymbol{z} \end{bmatrix}, \tag{23}$$

omitting the intermediate step of decoding to a virtual loudspeaker array (compare Fig. 2). It is more efficient to directly determine the filter weights subject to a set of given constraints, so that the resulting signal complies with known objective specifications. The filter weights for the left and right estimated binaural signal, $\boldsymbol{w}_\mathsf{L}$ and $\boldsymbol{w}_\mathsf{R}$, are the SH domain representation of a complex-valued function on the sphere that is supposed to

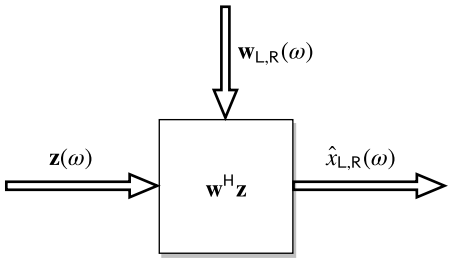


Figure 2: Binaural rendering filter.

assign the correct gain and delay to signals from any direction Ω. For the least squares decoder, we will come to the result that this function is the HRTF itself.

The goal is to minimize an objective function J that describes the error between between the estimated signal $\hat{\boldsymbol{x}}$ and a desired signal $\boldsymbol{x}$

$$\boldsymbol{w}^*(\omega) = \arg\min_{\boldsymbol{w}} J(\omega, \Omega) \tag{24}$$

For a plane wave $s(\omega)$ from direction Ω_0 at the input, encoded in N-th order Ambisonics

$$\boldsymbol{z}(\omega) = s(\omega)\, \boldsymbol{y}_N(\Omega_0), \tag{25}$$

the estimated signal can be expressed by

$$\hat{\boldsymbol{x}}(\omega) = s(\omega) \begin{bmatrix} \boldsymbol{w}_{\mathsf{L}}^{\mathsf{H}}(\omega)\boldsymbol{y}_N(\Omega_0) \\ \boldsymbol{w}_{\mathsf{R}}^{\mathsf{H}}(\omega)\boldsymbol{y}_N(\Omega_0) \end{bmatrix}. \tag{26}$$

The desired signal consists of the HRTF-filtered plane wave

$$\boldsymbol{x}(\omega) = s(\omega) \begin{bmatrix} H_{\mathsf{L}}(\omega, \Omega_0) \\ H_{\mathsf{R}}(\omega, \Omega_0) \end{bmatrix}. \tag{27}$$

Assuming that left and right HRTFs are symmetric, and dropping the frequency index, the cost function that minimizes the least squares error between the estimated and the desired signal for all directions on the sphere can be written as

$$\mathcal{J} = \int_{\Omega \in \mathcal{S}^2} |\boldsymbol{w}^{\mathsf{H}} \boldsymbol{y}(\Omega) - H(\Omega)|^2 \mathrm{d}\Omega, \tag{28}$$

and the filter weights that minimize the above expression are optimal in the lest-squares sense:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathcal{K}} \int_{\Omega \in \mathcal{S}^2} |\boldsymbol{w}^{\mathsf{H}} \boldsymbol{y}(\Omega) - H(\Omega)|^2 \mathrm{d}\Omega \tag{29}$$

In this equation, $\mathcal{K}$ is the domain over which $\boldsymbol{w}$ is optimized. In case of an unconstrained problem, $\mathcal{K} = \mathbb{C}^J$. If the problem is discretized, i.e., the sphere is sampled at a dense set of directions $\mathcal{M}$, the approximation of the above integral is given by

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathcal{K}}{\arg\min} \sum_{\Omega_j\in\mathcal{M}} |\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega_j) - H(\Omega_j)|^2 \equiv \underset{\boldsymbol{w}\in\mathcal{K}}{\arg\min} \|\boldsymbol{Y}_{\mathcal{M}N}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}}\|_2^2, \tag{30}$$

where $\boldsymbol{Y}_{\mathcal{M}N} = [\boldsymbol{y}_N(\Omega_j)]_{\Omega_j\in\mathcal{M}}$ and $\boldsymbol{h}_{\mathcal{M}} = [H(\Omega_j)]_{\Omega_j\in\mathcal{M}}$. The solution to problems of this kind can be found by multiplication with a pseudo-inverse. In case of the unconstrained LS problem, the solution is

$$\boldsymbol{w}^*_{LS} = \boldsymbol{Y}^{\dagger}_{\mathcal{M}N}\boldsymbol{h}_{\mathcal{M}} = \boldsymbol{Y}_{\mathcal{M}N}\boldsymbol{h}_{\mathcal{M}} = \mathcal{SHT}^{\mathcal{M}}_N\{\boldsymbol{h}_{\mathcal{M}}\}, \tag{31}$$

which can be interpreted as the spherical harmonic transform of the HRTF set.

Unfortunately, the unconstrained least squares decoder introduces colorations (spectral roll-off towards higher frequencies) if the input signal has a low order – which is the most common case. The least squares decoder tries to approximate the HRTFs for the whole sphere by means of a weighted linear combination of orthogonal basis functions (spherical harmonics). The weights are the spherical harmonic transform coefficients of the HRTF set. In case of an input signal with low order, less basis functions are available for the linear combination and the HRTFs can only be approximated with less spatial detail. But due to the off-center location of the ears, phase fluctuations between neighboring source positions lead to increasing spatial complexity with frequency. Thus, higher orders substantially contribute to the total energy.

If the SH transform in (31) is truncated at order N, the energy contained in the higher orders is dropped, leading to a roll-off towards higher frequencies. Consider the HRTF coefficients

$$\boldsymbol{h} = [H(\Omega_j)]_{j=1,\ldots,J} = \mathcal{ISHT}\{\mathring{\boldsymbol{h}}_M\} = \boldsymbol{Y}_M^{\mathsf{T}}\,\mathring{\boldsymbol{h}}_M \tag{32}$$

as the inverse SHT of the SH domain representation of the same HRTF with sufficiently high order M to represent the highest contained spatial frequency. The order-N SHT in (31) can be regarded as the projection of $\boldsymbol{Y}_M^{\mathsf{T}}\mathring{\boldsymbol{h}}_M$ onto $\boldsymbol{Y}_N$. Due to the order mismatch $M \gg N$, higher-order basis functions of $\boldsymbol{Y}_M^{\mathsf{T}}$, i.e., the orders $N+1,\ldots,M$, are projected onto the nullspace of $\boldsymbol{Y}_N$ (and vanish).

The least squares decoder in (26) can be interpreted as a virtual loudspeaker decoder which employs all HRTF measurement points as virtual loudspeakers. Based on this interpretation, Fig. 3 illuminates the problem from a different perspective. Consider the head of a listener placed at the coordinate origin, facing positive x direction, with the inter-aural axis lying on the y axis, and a single plane wave arriving from a point source in positive y direction. The point source corresponds to a spatial Dirac impulse in the ideal case. Order-truncated reproduction of this point source in Ambisonics entails the degeneration of the spatial Dirac impulse driving the virtual loudspeakers into a main lobe and various side lobes. The main lobe can be modeled as an array of isophasic

monopoles (corresponding to the HRTF measurement positions) located at the cap of a sphere around the origin.

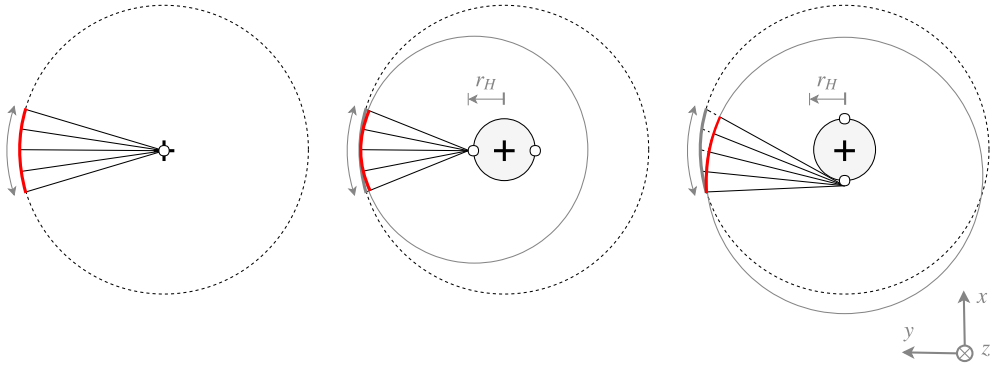


Figure 3: Isophasic monopoles driven by a degenerated beam lead to destructive interference in case of eccentric receivers, cf. [7].

Assume first that the head radius r_H approaches zero (Fig. 3, *left*), so that both ears would be located in the coordinate origin. Any two monopoles located on a sphere around the origin would have the same distance to the center, and therefore sound would arrive in-phase at the ears and interfere constructively. If the ears are moved off-center by $r_H \gg 0$ in $\pm y$ direction (Fig. 3, *center*), the path lengths from each of the monopoles to the ears vary slightly, leading to phase differences. Phase differences increase towards high frequencies and lead to destructive interference at multiple frequencies. The increased density of cancellations at high frequencies is perceived as a low-pass effect. If the point source is moved towards the frontal direction, the phase differences further increase, and so does the roll-off towards higher frequencies.

2.3.3 Binaural Modal Order Reduction

The authors in [7] propose to use only a reduced subset of the available $|\mathcal{M}|$ measurement points. This reduces the number of sources in the above model. If less sources interfere at the receivers, the attenuation of high frequencies is reduced while the total energy is retained. This is equivalent to spherical sub-sampling. While the maximum contained spatial frequency, which is determined by the spatial complexity of the HRTFs, remains constant, the number of available support points is reduced and higher-order components are thus aliased into lower orders. Even though the method yields good results in a listening test, it should be considered that, as the directional distribution of energy in the aliased components is not related to the original source distribution in a meaningful way anymore, localization is likely to be adversely affected.

2.3.4 Diffuse Field Equalization

The authors in [8] propose a global diffuse field equalization filter which is equivalent to restricting the optimization domain to [9]

$$\mathcal{K} = \left\{ \boldsymbol{w} \in \mathbb{C}^J : \boldsymbol{w}^{\mathsf{H}} \boldsymbol{w} = \int_{\Omega} |H(\Omega)|^2 = \|\boldsymbol{h}_{\mathcal{M}}\|_2^2 \right\}, \tag{33}$$

i.e., to require the norm of the rendering weights to be equal to the total norm of the HRTFs for all directions and frequencies. The solution to the constrained problem is

$$\boldsymbol{w}_{LS\,eq} = \frac{\|\boldsymbol{h}_{\mathcal{M}}\|_2}{\|\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}} \boldsymbol{h}_{\mathcal{M}}\|_2} \boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}} \boldsymbol{h}_{\mathcal{M}}. \tag{34}$$

While reducing the low-pass effect on average, direction-dependent distortions remain for low input orders because the modal order of each HRTF coefficient depends on the respective direction.

2.3.5 ITD Equalization

Fig. 4(a) shows the average energy contained in each order of an SH-transformed HRTF set over frequency. As can be seen from the figure, higher orders substantially contribute to the total energy. The increase in spatial complexity of HRTFs at high frequencies is due to rapid phase changes at high frequencies. In a recent contribution [10], it was shown that removing the linear phase, i.e., the phase differences between the left and right ear, reduces the amount of energy contained in higher order components. Consequently, HRTFs can be represented with a lower order in the SH domain, as shown in Fig. 4(b), without causing the low-pass effect that arises if the energy contained in higher orders is dropped. It has been shown that inter-aural phase differences (ITDs) do not contribute to localization at high frequencies. Hence, ITD equalization can be applied with no perceptual ramifications if the phase of lower frequencies remains unchanged.

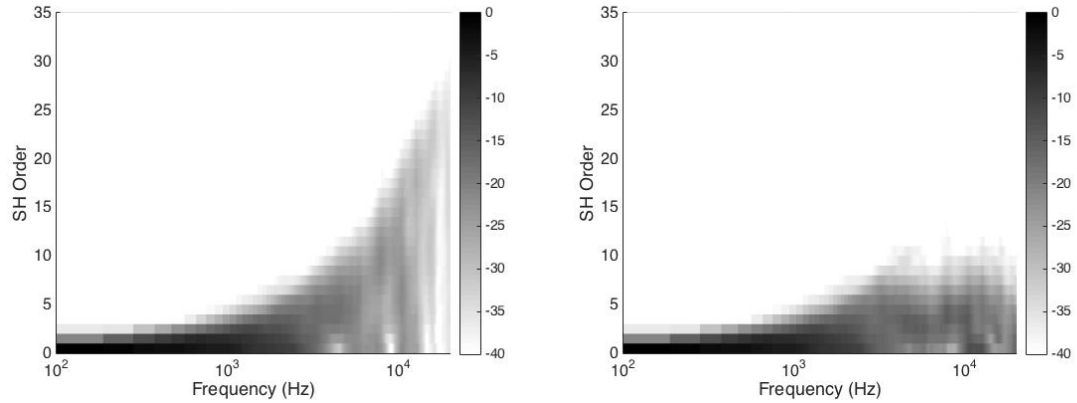
Frequency-dependent equalization can be achieved by multiplication of the original HRTFs with an an all-pass filter

$$\underline{H}(\omega, \Omega) = H(\omega, \Omega) A(\omega, \Omega) \tag{35}$$

with the frequency response

$$A(\omega, \Omega) = \begin{cases} 1 & \text{if } \omega < \omega_c \\ e^{-i(\omega - \omega_c)\tau(\Omega)} & \text{if } m < 0 \end{cases}, \tag{36}$$

where i is the imaginary unit, and $\omega_c = 2\pi \cdot 1.5\,\text{kHz}$ is the cut-on frequency above which the linear phase is compensated. $\tau_{\mathsf{L}}(\Omega)$ and $\tau_{\mathsf{R}}(\Omega)$ are the time difference between



(a) HRTF prior to modification (b) Modified above 3 kHz

Figure 4: Relative modal energy distribution of an HRTF set over frequency and impact of ITD equalization (HRTF of Neuman KU-100, cf. [11]).

the center and left or right ear, respectively. τ can be estimated from HRTF sets, but approximation on the basis of a spherical head model

$$\tau_{\mathsf{R}}(\Omega) \equiv \tau_{\mathsf{R}}(\vartheta, \varphi) = \cos\vartheta \sin\varphi r_H c^{-1},\ \tau_{\mathsf{R}}(\Omega) = -\tau_{\mathsf{L}}(\Omega), \tag{37}$$

where c is the speed of sound, and the head radius is set to $r_H = 8.5$ cm, works sufficiently well.

From the perspective of Fig. 3, the receivers are virtually shifted to the center for the upper freqency range. Since sound from all monopoles arrives in-phase destructive interference is prevented.

2.3.6 Magnitude Least Squares Decoder

In the previous section it was shown that subtracting the linear phase $\varphi_l(\Omega)$ at higher frequencies, i.e., setting the phase of the HRTFs to $\gamma(\Omega) := \angle H(\Omega) - \varphi_l(\Omega)$, reduces the spatial complexity of HRTFs and, therefore, the required SH expansion order for accurate reproduction. However, the residual phase does not necessarily lead to the lowest spatial complexity.

If the HRTFs are decomposed into the product of a matrix $\boldsymbol{M} = \mathrm{diag}|\boldsymbol{h}_{\mathcal{M}}|$ containing the absolute values and a phase vector $\boldsymbol{p} = [e^{i\gamma\Omega}]_{\Omega\in\mathcal{M}}$, where $\gamma(\Omega) = \angle H(\Omega)$, the cost function (30) can be written as [9]

$$\min_{\boldsymbol{w}\in\mathcal{K}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2. \tag{38}$$

After this decomposition it is possible to perform joint minimization over $\boldsymbol{w}$ and $\boldsymbol{p}$

$$\min_{\boldsymbol{w},\boldsymbol{p}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2 \tag{39}$$

$$\text{s.t. } |p_j| = 1 \; \forall j = 1, \ldots, |\mathcal{M}|, \tag{40}$$

i.e., determine the optimal filter coefficients and the optimal HRTF phase modification at the same time. Since the phase constraint is equivalent to ignoring the phase error and just minimizing over the magnitudes of the HRTFs, problem can be reformulated as $\min_{\boldsymbol{w}} \||\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}|\|_2^2$. Similar to the phase modification in the previous section, the phase modification is only applied above the cut-on frequency ω_c:

$$\boldsymbol{w}_{MagLS}^{*} = \arg\min_{\boldsymbol{w}} \begin{cases} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}}\|_2^2, & \omega \leq \omega_c \\ \||\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}|\|_2^2, & \omega > \omega_c \end{cases} \tag{41}$$

2.4 Methods for the Enhancement of the Spatial Resolution

The spatial resolution of first-order Ambisonics (FOA), and of binaural signals derived from it, is often insufficient. Various methods exist that try to enhance the resolution, of which a selection will be presented in the following.

2.4.1 Directional Audio Coding

Directional Audio Coding (DirAC) [12] is a parametric time-frequency domain method for the reproduction and enhancement of spatial audio for variable reproduction systems. In [13], it was combined with a method for generation of binaural output. The underlying assumption of DirAC is that the auditory system is limited to decoding only one cue for source direction, and one for diffuseness at a single time instant in each critical band. The directional cue is obtained from inter-aural time differences, inter-aural level differences and spectral cues, whereas the diffuseness cue refers to inter-aural coherence. It is assumed that the original signal is resynthesized perceptively correct if those very cues are faithfully reproduced within the respective frequency bands.

DirAC is subdivided into an analysis and a synthesis stage, which makes it applicable for transmission and storage of spatial audio. The analysis stage derives a parametric time-frequency representation of the B-Format input signal, consisting of the estimated direction of arrival Ω, the estimated diffuseness coefficient ψ, and one or multiple audio signals extracted from the input (depending on whether the focus lies on bit rate reduction for transmission or high quality reproduction). The direction of arrival (DOA) is defined as the negative intensity vector

$$\boldsymbol{D} = -\boldsymbol{I} = -p\boldsymbol{u}, \tag{42}$$

where p is the omnidirectional sound pressure and $\boldsymbol{u}$ is the particle velocity vector. If the sound pressure p is replaced by the omnidirectional w signal, the particle velocity vector

$\boldsymbol{u}$ is replaced by a vector containing the three figure-of-eight signals x, y, z, the result

$$\boldsymbol{D} \propto -\mathrm{Avg}_{\tau_1} \left\{ w^* \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right\} \tag{43}$$

is proportional to $\boldsymbol{D}$. The temporal average over a time interval τ_1, denoted by $\mathrm{Avg}_{\tau_1}\{\cdot\}$, is introduced in order to reduce artifacts caused by fast directional variations of the intensity vector. The temporal average is implemented as a weighted moving average, i.e., the samples in the interval $[t, t+\tau_1]$ are multiplied by a window function W_1 and then summed up. Since the length of $\boldsymbol{D}$ is irrelevant for the application,

$$\boldsymbol{D}' := -\sum_{m=t_1}^{t_1'} w(n+m)\boldsymbol{v}(n+m)W_1(m) \tag{44}$$

can be used as an estimate for the DOA. The diffuseness coefficient ψ is defined as one minus the relative fraction of direct sound, which is the ratio between the fraction of sound energy impinging from the DOA and the total sound energy:

$$\psi = 1 - \frac{\|\mathbb{E}\{\boldsymbol{I}\}\|}{c\mathbb{E}\{E\}} \approx 1 - \frac{\left\|\mathrm{Avg}_{\tau_2}\{\boldsymbol{I}\}\right\|}{c\mathrm{Avg}_{\tau_2}\{E\}},\ \psi \in [0,1], \tag{45}$$

where c is the speed of sound, $\mathrm{Avg}_{\tau_2}\{\cdot\}$ denotes the temporal average over the time interval $[t, t+\tau_2]$, and

$$E = \frac{\rho_0}{2}\left(\frac{p^2}{Z_0^2} + u^2\right) \tag{46}$$

is the instantaneous sound energy. The diffuseness for B-Format signals can be computed with

$$\psi \approx 1 - \frac{\sqrt{2}\|\sum_{m=t_2}^{t_2'} w(n+m)\boldsymbol{v}(n+m)W_2(m)\|}{\sum_{m=t_2}^{t_2'}\left(|w(n+m)|^2 + \frac{1}{2}|\boldsymbol{v}(n+m)|^2\right)W_2(m)}. \tag{47}$$

The synthesis stage receives the DOA and diffuseness parameters from the analysis stage, and also either the omnidirectional w signal or virtual microphone signals, for the low-bandwidth or high-quality variant, respectively. The synthesis stage tries to resynthesize the sound field as the superposition of a directional and a diffuse sound component. Thus, the audio signals are split up into a diffuse and a direct stream which are then multiplied by $\sqrt{\psi}$ or $\sqrt{1-\psi}$, respectively. The diffuse stream is then decorrelated and fed to the loudspeakers. Decorrelation is needed to reduce the coherence between the loudspeaker signals in order to achieve the impression of a diffuse sound field. The direct stream is used to create point-like virtual sources at positions corresponding to the DOA.

The omnidirectional channel contains a mixture of direct and diffuse sound. It is desirable to increase the amount of direct sound in the point source signals, while, at the same time, to reduce the coherence between the diffuse loudspeaker signals. This is achieved by extracting the signals of virtual microphones pointing in the direction of each loudspeaker. In principle, different reproduction techniques can be applied, such as Vector Based Amplitude Panning (VBAP), or Ambisonics. In the case of VBAP, this corresponds to three virtual microphone signals pointing to the loudspeaker triangle currently playing (direct component) and one virtual microphone signal per loudspeaker (diffuse component). Another benefit of using virtual microphones is that the diffuse signals only have to be mildly decorrelated (decorrelation is a major source of artifacts). The creation of virtual microphones, however, requires transmission of the full B-Format signal instead only transmitting the w component.

For binaural reproduction, as proposed in [13], the loudspeakers are replaced by virtual loudspeakers. Each virtual loudspeaker signal is convolved with the HRIR corresponding to the loudspeaker position and then summed up for both ears. Using a fixed setup of virtual loudspeakers, there is no need for HRTF interpolation in real-time. Furthermore, sound scene rotation according to head movements of the listener (head-tracking) can be achieved by transformation of the direction metadata and the virtual microphone positions.

2.4.2 High Angular Resolution Planewave Expansion

High Angular Resolution Planewave Expansion, or HARPEX, is a time-frequency domain method to enhance B-Format signals. In [14], it was proposed for decoding to general surround loudspeaker configurations, and can therefore be combined with arbitrary panning methods, such as pairwise panning, Ambisonics, or wave field synthesis. In [15], the authors extend the method to generate binaural output signals.

HARPEX is based on source separation by plane wave decomposition and subsequent higher-order encoding of the extracted plane waves. A B-Format input vector $\boldsymbol{X}$ is decomposed into the matrices $\boldsymbol{V}$ and $\boldsymbol{A}$,

$$\underbrace{\begin{bmatrix} \sqrt{2}w \\ x \\ y \\ z \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} 1 & 1 \\ x_1 & x_2 \\ y_1 & y_2 \\ z_1 & z_2 \end{bmatrix}}_{\boldsymbol{V}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{\boldsymbol{A}}, \tag{48}$$

where the lower 3×2 matrix of $\boldsymbol{V}$ contains the incidence directions, and $\boldsymbol{A}$ contains the complex amplitudes of two plane waves. $\boldsymbol{X}$ is split into its real and imaginary part and then decomposed by means of a QR decomposition. The QR decomposition finds the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ which satisfy

$$[\mathfrak{Re}(\boldsymbol{X})\ \mathfrak{Im}(\boldsymbol{X})] = \boldsymbol{Q}\boldsymbol{R}, \tag{49}$$

where the columns of the 4×2 matrix $\boldsymbol{Q}$ form an orthonormal basis for the column space of the decomposed matrix, and $\boldsymbol{R}$ contains the linear combination coefficients. $\boldsymbol{Q}$ is then transformed into $\boldsymbol{V}$ with respect to the conditions that (i) the square of the top element of $\boldsymbol{V}$ is equal to the squared norm of the bottom three elements, and (ii) the top element is equal to 1. A matrix

$$\boldsymbol{D} = \boldsymbol{Q}\boldsymbol{C} \tag{50}$$

that satisfies (i) can be found by multiplying the matrix $\boldsymbol{Q}$ by a 'mixing matrix' $\boldsymbol{C}$, where $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = Q_{11} \begin{bmatrix} 1 & 1 \\ -b & b \end{bmatrix} + Q_{12} \begin{bmatrix} b & -b \\ 1 & 1 \end{bmatrix}, \tag{51}$$

$$b = \sqrt{2(Q_{11}^2 + Q_{12}^2) - 1}. \tag{52}$$

In order to satisfy (ii), the columns of $\boldsymbol{D}$ are divided by their top elements

$$\boldsymbol{V} = \boldsymbol{D} \begin{bmatrix} D_{11}^{-1} & 0 \\ 0 & D_{12}^{-1} \end{bmatrix}. \tag{53}$$

The plane wave amplitudes $\boldsymbol{A}$ are not required for the further processing steps, but can be determined from $\boldsymbol{D}$, $\boldsymbol{C}$, and $\boldsymbol{R}$.

If $Q_{11}^2 + Q_{12}^2 < \frac{1}{2}$, the solution to (52) is not defined and the decomposition does not exist. In these cases, an alternative method must be employed. The authors claim that the choice of the fall-back method does not make any perceptible difference because it affects only a small fraction of all samples.

The basic decoder extracts two plane waves and uses the directions of arrival of the extracted plane waves as virtual loudspeaker positions. An HRTF set is evaluated at the directions of arrival, which may require interpolation, and the HRTF coefficients are then multiplied by the complex amplitude of the respective plane wave. The number of extracted plane waves, which is two, is derived from the available degrees of freedom of the input vector $\boldsymbol{X}$, which is eight: real and imaginary part of the 4 B-Format signals. The authors further argue that discriminating of two sources per time-frequency bin is already at the limit of the capability of the auditory system.

However, the basic decoder introduces artifacts because of sharp edges in the transfer function of the resulting system. Smoothing can only be applied to the transfer function if it is known. The transfer function can be obtained by multiplying the input-to-plane-waves decoder by the plane waves-to-binaural decoder (the so-called output mode vectors), where the input-to-plane-waves decoder is the inverse of a matrix containing the plain wave representation (the so-called input mode vectors), similar to $\boldsymbol{V}$. The result is a decoding matrix that decodes the input vector to the binaural output signals. In order to be able to compute the inverse of the input mode matrix, and to ensure

it is well-conditioned, two more plane waves are added to the input mode vectors at complementary directions.

Another source of artifacts are DOA estimation errors. These errors lead to phase noise. The phase difference between the correct HRTF coefficient and the one selected at the noisy DOA increases with frequency. While phase noise is audible over the whole frequency range, phase, or more specifically, inter-aural time differences are only evaluated at low and mid frequencies. Since phase noise in the low range is tolerable, the phase for high frequencies is replaced by another less accurate, but also less noisy estimate. The new estimate is obtained by summing up the group delays over frequency. It may be affected by offset errors, which are, however, inaudible at high frequencies.

2.4.3 Sparse Plane Wave Decomposition

In [16], a method was proposed that tries to estimate higher order Ambisonic signals from an Ambisonic input signal via convolution with a filter matrix. The filter matrix is obtained using sparse plane wave decomposition. Plane wave decomposition tries to explain acoustic observations as the superposition of plane wave signals. The plane wave decomposition of an order-L Ambisonic input signal vector $\boldsymbol{b} = [b_0^0 \; b_0^{-1} \; \cdots \; b_L^L]^\mathsf{T}$ in the time/frequency domain is given by

$$\boldsymbol{b}(t,k) = \boldsymbol{Y}\boldsymbol{s}(t,k), \tag{54}$$

where k is the index of the frequency band, t is the index of the time frame, $\boldsymbol{s}$ is a vector containing the plane wave source signals corresponding to the directions $\Omega_1 \dots \Omega_P$, and

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}(\Omega_1) & \boldsymbol{y}(\Omega_2) & \cdots & \boldsymbol{y}(\Omega_P) \end{bmatrix}, \tag{55}$$

which the authors refer to as a *spatial dictionary* for the plane-wave decomposition, is a matrix of spherical harmonic vectors evaluated for a set of P points regularly distributed on the sphere with $\boldsymbol{y}(\Omega_i) = [Y_0^0(\Omega_i)\, Y_0^{-1}(\Omega_i) \cdots Y_L^L(\Omega_i)]^\mathsf{T}$. The size of dictionary must be much greater than the number of Ambisonics signals $P \gg (L+1)^2$, and $P \geq (L'+1)^2$, where $L' > L$ is the order of the Ambisonic output signal. Since (54) is an underdetermined linear system, infinitely many solutions exist.

Solving for $\boldsymbol{s}$ using the Moore-Penrose pseudoinverse leads to the least-norm solution which, however, is not necessarily the best choice in a perceptual sense. It distributes the plane wave energy evenly over the sphere, but since in most cases sound sources playing at the same time and in the same frequency band are expected to be discrete in space and their number is expected to be small, it is makes sense to solve for the solution most sparsely distributed in space instead. This kind of problem is known from compressed sensing (CS) techniques, and a common approach would be to minimize the ℓ_1 norm of $\boldsymbol{s}$ subject to $\boldsymbol{Y}\boldsymbol{s}(n,k) = \boldsymbol{b}(n,k)$. The ℓ_1 norm

$$\|\boldsymbol{s}\|_1 := \sum_{i=1}^{P} |s_i| \tag{56}$$

is the sum of the absolute values of each element of $\boldsymbol{s}$. Minimizing the ℓ_1 norm leads to a convex optimization problem that needs to be solved for each time/frequency sample separately. But discontinuities between neighboring bins can lead to artifacts. In order to reduce these artifacts, multiple time slices of the input signal must be considered, which transforms the problem into a multiple-measurement vector type optimization problem:

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{S}(n,k)\|_{12} && (57)\\ \text{subject to} \quad & \boldsymbol{Y}\boldsymbol{S}(n,k) = \boldsymbol{B}(n,k), && (58) \end{aligned}$$

where

$$\boldsymbol{B}(n,k) = [\boldsymbol{b}(n\tau+1,k)\,\boldsymbol{b}(n\tau+2,k)\,\cdots\boldsymbol{b}(n\tau+T,k)]\; n \in \mathbb{N} \tag{59}$$

and

$$\boldsymbol{S}(n,k) = [\boldsymbol{s}(n\tau+1,k)\,\boldsymbol{s}(n\tau+2,k)\,\cdots\boldsymbol{s}(n\tau+T,k)]. \tag{60}$$

$\|\cdot\|_{12}$ denotes the ℓ_1 norm of the row-wise ℓ_2 norm of a matrix, i.e.,

$$\|\boldsymbol{A}\|_{12} := \sum_i \sqrt{\sum_j |a_{ij}|^2}. \tag{61}$$

If the problem is considered as finding a *demixing matrix* $\boldsymbol{D}$ with

$$\boldsymbol{S}(n,k) = \boldsymbol{D}(n,k)\boldsymbol{B}(n,k) \tag{62}$$

which computes the plane wave vectors from the Ambisonic input vectors, it can be reformulated as

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{D}(n,k)\boldsymbol{B}(n,k)\|_{12} && (63)\\ \text{subject to} \quad & \boldsymbol{Y}\boldsymbol{D}(n,k) = \boldsymbol{I}, && (64) \end{aligned}$$

and can be solved using the Iteratively Re-weighted Least Squares algorithm (IRLS). The successive measurement vectors are not necessarily linearly independent. Thus, solving in the subspace spanned by the first $(L+1)^2$ singular vectors can effectively reduce the complexity of the problem. An upscaling matrix

$$\boldsymbol{U}(n,k) = (1-\alpha)\boldsymbol{U}(n-1,k) + \alpha\boldsymbol{Y}'\boldsymbol{D}(n,k) \tag{65}$$

is computed from the demixing matrix, where the forgetting factor $\alpha \in [0,1]$ is used for smoothing, and $\boldsymbol{Y}'$ is a matrix of SH vectors defined identically to (55), but expanded

up to order L'. The increased-order Ambisonic output signal can be obtained via multiplication by $\boldsymbol{U}$ in the frequency domain,

$$\boldsymbol{B}'(n,k) = \boldsymbol{U}(n,k)\boldsymbol{B}(n,k), \tag{66}$$

or via time-frame-wise convolution in the time domain,

$$\boldsymbol{b}'^{(n)}(t) = (\boldsymbol{U}^{(n)} * \boldsymbol{b}^{(n)})(t), \tag{67}$$

followed by overlap and add.

3 Single-Parametric Binaural Rendering

In Section 2.3.2, the least-squares method for rendering Ambisonics to binaural signals was explained. For input signals with a low order, the unconstrained least squares solution introduces severe artifacts, and the attained resolution is low. These problems arise from the difficulty to approximate the complex directivity pattern of unmodified HRTFs adequately by a linear combination of lower-order spherical harmonics. The performance of the least-squares decoder can be improved with various approaches, e.g., global diffuse field equalization or reduction of the spatial complexity of the HRTFs by phase modification. Note that these approaches are unable to increase the resolution above the resolution of the input signal.

However, different methods exist which try to increase the spatial resolution. In contrast to the least-squares decoder, which is a time-invariant linear filter and independent of the input, the methods presented in the previous section are parametric and time-variant, and aim at the resynthesis of the input signal in order to gain a higher resolution. Ultimately, resynthesis is either based on the explanation of each observation (frame) by a number of plane waves (HARPEX, Sparse Plane Wave Decomposition), or a single plane wave and diffuse sound (DirAC) and requires the extraction of parameters from the input signal (direction of arrival, diffuseness, number and distribution of plane waves)

This section presents a recently proposed method [17] that uses a parametric linear filtering approach. In addition to minimizing the least-squares error for the whole sphere, two constraints are imposed upon this optimization problem to yield accurate reproduction of direct sound while preserving the statistical characteristics of diffuse sound. To begin with, the underlying signal model will be explained. Subsequently, after formulating the constraints, the approach to solving this optimization problem will be explained.

3.1 Signal Model

We model a sound field in the Ambisonics domain with limited order P by the sum of a direct sound component $\boldsymbol{z}_d$, which is the superposition of D source signals s_δ arriving from directions Ω_δ, and a diffuse component $\boldsymbol{z}_n$. The frequency-domain input signal $\boldsymbol{z}(\omega, t)$ at a given time instant t is given by

$$\begin{aligned} \boldsymbol{z}(\omega,t) &= \boldsymbol{z}_d(\omega,t) + \boldsymbol{z}_n(\omega,t) \\ &= \sum_{\delta=1}^{D} \boldsymbol{z}_\delta(\omega,t) + \boldsymbol{z}_n(\omega,t) \\ &= \sum_{\delta=1}^{D} s_\delta(\omega,t)\, \boldsymbol{y}_P(\Omega_\delta) + \boldsymbol{z}_n(\omega,t), \end{aligned} \tag{68}$$

where $\boldsymbol{z}(\omega,t) = [z_0^0(\omega,t) \cdots z_P^P(\omega,t)]^\mathsf{T}$ holds the Ambisonic signals, $s_\delta(\omega,t)$ is the direct sound signal of the δ-th plane wave, and $\boldsymbol{y}_P(\Omega_\delta) = [Y_0^0(\Omega_\delta) \cdots Y_P^P(\Omega_\delta)]^\mathsf{T}$ is a vector

of real-valued spherical harmonics up to order P evaluated at Ω_δ. If we assume only a single plane wave $s_0(\omega, t)$ arriving from Ω_0 to contribute to the sound field at a given time instant and frequency index, (68) simplifies to

$$\boldsymbol{z}(\omega, t) = s_0(\omega, t)\, \boldsymbol{y}_P(\Omega_0) + \boldsymbol{z}_n(\omega, t). \tag{69}$$

This assumption is already familiar from Directional Audio Coding, cf. Section 2.4.1. Similarly, we model the binaural signal as the superposition of a direct and a diffuse binaural signal

$$\begin{aligned} \boldsymbol{x}(\omega, t) &= \boldsymbol{x}_d(\omega, t) + \boldsymbol{x}_n(\omega, t) \\ &= \sum_{\delta=1}^{D} \boldsymbol{x}_\delta(\omega, t) + \boldsymbol{x}_n(\omega, t) \\ &= \sum_{\delta=1}^{D} s_\delta(\omega, t)\, \boldsymbol{h}(\Omega_\delta, \omega) + \boldsymbol{x}_n(\omega, t), \end{aligned} \tag{70}$$

where $\boldsymbol{x}(\omega, t) = [x_\mathsf{L}(\omega, t)\ x_\mathsf{R}(\omega, t)]^\mathsf{T}$ contains the signals for the left and right ear, $\boldsymbol{x}_\delta$ is δ-th direct signal arriving from direction Ω_δ, $\boldsymbol{x}_n$ is the diffuse binaural component, and

$$\boldsymbol{h}(\Omega_\delta, \omega) = [h_\mathsf{L}(\Omega_\delta, \omega)\ h_\mathsf{R}(\Omega_\delta, \omega)]^\mathsf{T} \tag{71}$$

is the HRTF for each of the two ears evaluated at frequency ω and direction Ω_δ. Again, if we assume only a single source to be active at each time/frequency sample, the model can be simplified to

$$\boldsymbol{x}(\omega, t) = s_0(\omega, t)\, \boldsymbol{h}(\Omega_0, \omega) + \boldsymbol{x}_n(\omega, t). \tag{72}$$

3.2 Formulation of the Constraints

We estimate the binaural signal by means of a linear transformation

$$\hat{\boldsymbol{x}}(\omega, t) = \boldsymbol{W}(\Omega_0, \omega)\, \boldsymbol{z}(\omega, t), \tag{73}$$

where $\hat{\boldsymbol{x}}(\omega, t)$ is the estimated binaural signal, and the $2\times(P+1)^2$ matrix $\boldsymbol{W} = [\boldsymbol{w}_\mathsf{L}\ \boldsymbol{w}_\mathsf{R}]^\mathsf{H}$ is a complex weighting matrix. The vectors $\boldsymbol{w}_\mathsf{L}$ and $\boldsymbol{w}_\mathsf{R}$ are the weights of the linear combination of the Ambisonic input signal yielding signals for the left and right ear, respectively. Note that $\boldsymbol{W}$ depends on the direction of arrival, Ω_0. It can be regarded as the solution to a beamforming problem that meets the following conditions at the same time:

(1) Best possible localization of direct sound

(2) Unaltered impression of the diffuse sound field

If both conditions are met, we expect perceptual equivalence. A mathematical formulation of the problem is developed in the following. Note that the frequency and time indices are omitted for the sake of readability.

3.2.1 Linear Constraint

Best possible localization of the direct component is achieved if

$$\hat{\boldsymbol{x}}_d = \boldsymbol{x}_d, \tag{74}$$

i.e., the estimated direct component is equal to the product of the unknown direct source signal and the HRTF coefficient of the respective direction and frequency. In Section 2.3.5 it was shown that we can reduce the spatial complexity of an HRTF by ITD equalization for high frequencies without detrimental perceptual effects. We therefore define

$$\boldsymbol{x}_d := \underline{\boldsymbol{h}}(\Omega_0) s_0, \tag{75}$$

where the underline indicates that phase modification was applied. Inserting (75) and the estimated binaural signal

$$\hat{\boldsymbol{x}}_d = \boldsymbol{W}(\Omega_0)\boldsymbol{y}_P(\Omega_0) s_0 \tag{76}$$

into (74) leads to a linear constraint on $\boldsymbol{W}(\Omega_0)$:

$$\boldsymbol{W}(\Omega_0)\boldsymbol{y}_P(\Omega_0) = \underline{\boldsymbol{h}}(\Omega_0) \tag{77}$$

3.2.2 Covariance Constraint

In order to take the diffuse field into account, we need to find another constraint on the weighting matrix. For correct encoding of the diffuse component, the diffuse field energy and the interaural coherence of the sound field should not be altered. A sufficient stochastic description of the binaural signals is provided by the auto-covariance matrix of the desired signal, which holds the variances (signal power) of the left and right ear signals along the main diagonal, and their covariance (joint variability) as the top right and bottom left element. The covariance matrix, which is identical to the correlation matrix for zero-mean signals, has the form

$$\boldsymbol{\Phi}_x = \mathbb{E}\left\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right\} = \begin{bmatrix} \mathbb{E}\{x_{\mathsf{L}} x_{\mathsf{L}}^*\} & \mathbb{E}\{x_{\mathsf{L}} x_{\mathsf{R}}^*\} \\ \mathbb{E}\{x_{\mathsf{L}}^* x_{\mathsf{R}}\} & \mathbb{E}\{x_{\mathsf{R}} x_{\mathsf{R}}^*\} \end{bmatrix} = \begin{bmatrix} \sigma_{\mathsf{L}}^2 & r_{\mathsf{LR}}^* \\ r_{\mathsf{LR}} & \sigma_{\mathsf{R}}^2 \end{bmatrix}, \tag{78}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator, i.e., the average over multiple realizations of the same stochastic process. If the diffuse field is unaltered, the covariance matrices of the estimated and the desired signal must be identical:

$$\mathbb{E}\Big\{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\mathsf{H}}\Big\} = \mathbb{E}\Big\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\Big\} \tag{79}$$

The covariance matrix of the desired signal can be computed by substituting (70) into the definition

$$\begin{aligned}\mathbb{E}\Big\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\Big\} &= \mathbb{E}\Big\{(s_0\,\boldsymbol{h}(\Omega_0) + \boldsymbol{x}_n)(s_0\,\boldsymbol{h}(\Omega_0) + \boldsymbol{x}_n)^{\mathsf{H}}\Big\} \\ &= \mathbb{E}\Big\{s_0^2\boldsymbol{h}(\Omega_0)\boldsymbol{h}^{\mathsf{H}}(\Omega_0) + 2s_0\boldsymbol{h}(\Omega_0)\boldsymbol{x}_n^{\mathsf{H}} + \boldsymbol{x}_n\boldsymbol{x}_n^{\mathsf{H}}\Big\}.\end{aligned} \tag{80}$$

In the above equation, $\mathbb{E}\Big\{s_0^2\Big\} = \sigma_d^2$ is the variance of the direct signal, and $\mathbb{E}\Big\{\boldsymbol{x}_n\boldsymbol{x}_n^{\mathsf{H}}\Big\} = \sigma_n^2\boldsymbol{\Phi}_H$ is the product of the variance of the diffuse signal and the diffuse field covariance matrix of the HRTF set. After the term $2s_0\boldsymbol{h}(\Omega_0)\boldsymbol{x}_n^{\mathsf{H}}$ vanishes because the input signal is assumed to be uncorrelated with the diffuse signal, the covariance matrix finally becomes

$$\mathbb{E}\Big\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\Big\} = \sigma_d^2\boldsymbol{h}(\Omega_0)\boldsymbol{h}^{\mathsf{H}}(\Omega_0) + \sigma_n^2\boldsymbol{\Phi}_H. \tag{81}$$

$\boldsymbol{\Phi}_H$ is the diffuse field covariance matrix of the (unmodified) HRTF set, i.e., the spatial integral over the covariance matrix of the HRTF directional patterns, weighted with the probability of each incidence direction. A perfectly diffuse sound field is defined as the superposition of an infinite number of plane waves with unit amplitude and random phase. Thus, each direction has equal probability and $\boldsymbol{\Phi}_H$ is given by

$$\boldsymbol{\Phi}_H = \mathbb{E}\Big\{\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\Big\} = \int_{S^2}\boldsymbol{h}(\Omega)\boldsymbol{h}^{\mathsf{H}}(\Omega)\mathrm{d}\Omega. \tag{82}$$

This integral can be approximated by the discrete sum over directions Ω_j,

$$\boldsymbol{\Phi}_H \approx \sum_{j=1}^{J} g_j\boldsymbol{h}(\Omega_j)\boldsymbol{h}^{\mathsf{H}}(\Omega_j) = \boldsymbol{H}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{H}^{*}, \tag{83}$$

where $\boldsymbol{H} = [\boldsymbol{h}(\Omega_1)\cdots\boldsymbol{h}(\Omega_J)]^{\mathsf{T}}$. The diffuse field covariance matrix can also be expressed in the spherical Fourier domain with $\boldsymbol{H} = \mathcal{ISFT}\{\mathring{\boldsymbol{H}}_N\}$ by

$$\boldsymbol{\Phi}_H \approx \boldsymbol{H}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{H}^{*} = \mathring{\boldsymbol{H}}_N^{\mathsf{T}}\boldsymbol{Y}_N\boldsymbol{G}\boldsymbol{Y}_N^{\mathsf{T}}\mathring{\boldsymbol{H}}_N^{*} = \mathring{\boldsymbol{H}}_N^{\mathsf{T}}\mathring{\boldsymbol{H}}_N^{*}. \tag{84}$$

The covariance matrix of the estimated signal is given by

$$\boldsymbol{\Phi}_{\hat{\boldsymbol{x}}} = \mathbb{E}\Big\{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\mathsf{H}}\Big\} = \mathbb{E}\Big\{\boldsymbol{W}(\Omega_0)\boldsymbol{z}\boldsymbol{z}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}(\Omega_0)\Big\}. \tag{85}$$

After computing the expectation of $\boldsymbol{z}\boldsymbol{z}^{\mathsf{H}}$, which is

$$\mathbb{E}\Big\{\boldsymbol{z}\boldsymbol{z}^{\mathsf{H}}\Big\} = \sigma_d^2\boldsymbol{y}_P(\Omega_0)\boldsymbol{y}_P^{\mathsf{H}}(\Omega_0) + \sigma_n^2\boldsymbol{I}, \tag{86}$$

where σ_d^2 and σ_n^2 are the variances of the direct and diffuse signal, respectively, the covariance matrix of the estimated signal can be written as

$$\mathbb{E}\left\{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^{\mathsf{H}}\right\} = \sigma_d^2 \boldsymbol{W}(\Omega_0)\boldsymbol{y}_P(\Omega_0)\boldsymbol{y}_P^{\mathsf{H}}(\Omega_0)\boldsymbol{W}^{\mathsf{H}}(\Omega_0) + \sigma_n^2 \boldsymbol{W}(\Omega_0)\boldsymbol{W}^{\mathsf{H}}(\Omega_0) \tag{87}$$

Inserting the linear constraint (77) into the above equation, and inserting expressions for the covariance matrix of the estimated and the desired signals into (79) leads to a second, quadratic constraint

$$\boldsymbol{W}(\Omega_0)\boldsymbol{W}^{\mathsf{H}}(\Omega_0) = \boldsymbol{\Phi}_H, \tag{88}$$

which will be referred to as the covariance constraint in the following.

3.3 Problem Formulation

The problem can be formulated as a constrained optimization problem,

$$\boldsymbol{W}_0 = \arg\min_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}) \tag{89}$$

$$\text{subject to: } (1)\ \boldsymbol{W}\boldsymbol{y}_0 = g_l \underline{\boldsymbol{h}}_0 \tag{90}$$

$$(2)\ \boldsymbol{W}\boldsymbol{W}^{\mathsf{H}} = g_c^2 \boldsymbol{\Phi}_H, \tag{91}$$

where $\boldsymbol{W}_0$ is defined as the matrix $\boldsymbol{W}$ that minimizes a cost function $\mathcal{J}$ for the direction of arrival (DOA) Ω_0 in the least-squares sense subject to the linear (90) and the quadratic constraint (91) derived before. g_l and g_c are positive scalars that are ideally equal to one. In case the constraints can not be exactly met, the constraints can be weighted with $\alpha(g_l - 1)^2 + (1-\alpha)(g_c - 1)^2 = \min.$ by the choice of $\alpha \in [0, 1]$.

We define the cost function as the least squares error for all directions on the sphere:

$$\mathcal{J}(\boldsymbol{W}) = \int_{\mathcal{S}^2} d(\Omega) \|\boldsymbol{W}\boldsymbol{y}_P(\Omega) - \underline{\boldsymbol{h}}(\Omega)\|_2^2 \mathrm{d}\Omega, \tag{92}$$

The cost function is identical to the cost function for the least-squares decoder in (28), except for the optional factor $d(\Omega)$ that can be used to put more weight on certain directions. If the cost function is constrained by (90) and (91), the weights optimized for a certain DOA are the matrix that minimizes the least squares error for all directions on the sphere while exactly meeting the linear constraint for plane waves arriving from the DOA and the quadratic constraint for the diffuse sound field.

3.4 Solution

All matrices $\boldsymbol{W}$ that satisfy the covariance constraint (91) can be expressed by

$$\boldsymbol{W} = g_c \boldsymbol{K} \boldsymbol{P}, \tag{93}$$

where $\boldsymbol{K}\boldsymbol{K}^{\mathsf{H}} = \boldsymbol{\Phi}_H$ and $\boldsymbol{P}$ is an arbitrary matrix such that $\boldsymbol{P}\boldsymbol{P}^{\mathsf{H}} = \boldsymbol{I}$, i.e., $\boldsymbol{P}$ is unitary. The matrix $\boldsymbol{K}$ can be computed with a suitable decomposition of $\boldsymbol{\Phi}_H$. Employing the eigen decomposition $\boldsymbol{\Phi}_H = \boldsymbol{U}_H \boldsymbol{S}_H \boldsymbol{U}_H^{\mathsf{H}}$ yields

$$\boldsymbol{K} = \boldsymbol{U}_H \boldsymbol{S}_H. \tag{94}$$

The matrix $\boldsymbol{P}$ is a means to parametrize the remaining degrees of freedom in order to meet the linear constraint as well. Note that due to the introduction of $\boldsymbol{P}$ we can modify the solution of the above equation, while the solution of

$$\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}} = g_c^2 \boldsymbol{K}\boldsymbol{P}\boldsymbol{P}^{\mathsf{H}}\boldsymbol{K}^{\mathsf{H}} = g_c^2 \boldsymbol{\Phi}_H \tag{95}$$

remains unchanged, so that the covariance constraint is still met.

Inserting (93) into the linear constraint (90) yields

$$\boldsymbol{P}\boldsymbol{y}_0 = g \boldsymbol{K}^{-1} \underline{\boldsymbol{h}}_0, \tag{96}$$

where $g = g_l / g_c$. The least-norm solution of the above equation is

$$\boldsymbol{P}_l = \frac{g}{q^2} \boldsymbol{K}^{-1} \underline{\boldsymbol{h}}_0 \boldsymbol{y}_0^{\mathsf{H}}. \tag{97}$$

However, the least-norm solution to (96) does not necessarily satisfy the covariance constraint. But, since we are allowed to add any expression to the solution that lies inside the nullspace of $\boldsymbol{y}_P$ without changing the solution to $\boldsymbol{P}\boldsymbol{y}_0$, we can write

$$\boldsymbol{P} = \boldsymbol{P}_l + \boldsymbol{R}\boldsymbol{N}_{\boldsymbol{y}}, \tag{98}$$

where $\boldsymbol{N}_y$ constitutes an orthonormal basis for the nullspace of $\boldsymbol{y}_0$, and $\boldsymbol{R}$ is an arbitrary matrix such that $\boldsymbol{P}$ is unitary. With $\boldsymbol{N}_y \boldsymbol{P}_l^{\mathsf{H}} = \boldsymbol{0}$ and $\boldsymbol{N}_y \boldsymbol{N}_y^{\mathsf{H}} = \boldsymbol{I}$, we have

$$\begin{aligned} \boldsymbol{I} &= \boldsymbol{P}\boldsymbol{P}^{\mathsf{H}} \\ &= (\boldsymbol{P}_l + \boldsymbol{R}\boldsymbol{N}_y)(\boldsymbol{P}_l + \boldsymbol{R}\boldsymbol{N}_y)^{\mathsf{H}} \\ &= \boldsymbol{P}_l \boldsymbol{P}_l^{\mathsf{H}} + \boldsymbol{R}\boldsymbol{R}^{\mathsf{H}}. \end{aligned} \tag{99}$$

If we define

$$\boldsymbol{R}\boldsymbol{R}^{\mathsf{H}} = \boldsymbol{C} := \boldsymbol{I} - \underbrace{\frac{g^2}{q^2}\boldsymbol{K}^{-1}\boldsymbol{h}_0\underline{\boldsymbol{h}}_0^{\mathsf{H}}\boldsymbol{K}^{-\mathsf{H}}}_{\boldsymbol{P}_l\boldsymbol{P}_l^{\mathsf{H}}}, \tag{100}$$

it becomes clear that this problem is similar to (91), and we can use the same ansatz,

$$\boldsymbol{R} := \boldsymbol{K}_C\boldsymbol{P}_C, \tag{101}$$

where $\boldsymbol{K}_C\boldsymbol{K}_C^{\mathsf{H}} = \boldsymbol{C}$ and $\boldsymbol{P}_C$ is an arbitrary $2 \times (P+1)^2 - 1$ matrix with orthonormal rows and $\boldsymbol{P}_C\boldsymbol{P}_C^{\mathsf{H}} = \boldsymbol{I}$. Hence, the solution for $\boldsymbol{W}_0$ is given by

$$\boldsymbol{W}_0 = g_c\boldsymbol{K}(\boldsymbol{P}_l + \boldsymbol{R}\boldsymbol{N}_y) = \frac{g_l}{q^2}\underline{\boldsymbol{h}}_0\boldsymbol{y}_0^{\mathsf{H}} + g_c\boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y, \tag{102}$$

where $\boldsymbol{K}_C = \boldsymbol{U}_C\boldsymbol{\Sigma}_C$, and $\boldsymbol{P}_C$ is yet to be determined. We reformulate the problem:

$$\boldsymbol{P}_C^* = \underset{\boldsymbol{P}_C}{\arg\min} \int_{\mathcal{S}^2} d(\Omega)\|\boldsymbol{W}\boldsymbol{y}_P(\Omega) - \underline{\boldsymbol{h}}(\Omega)\|_2^2 \mathrm{d}\Omega \tag{103}$$

$$\text{subject to: } (1)\ \boldsymbol{W} = \frac{g_l}{q^2}\underline{\boldsymbol{h}}_0\boldsymbol{y}_0^{\mathsf{H}} + g_c\boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y \tag{104}$$

$$(2)\ \boldsymbol{P}_C\boldsymbol{P}_C^{\mathsf{H}} = \boldsymbol{I} \tag{105}$$

we rewrite the problem in matrix form, i.e., approximate the integral numerically for a given set of points,

$$\boldsymbol{P}_C = \underset{\boldsymbol{P}_C}{\arg\min} \left\| \left(\boldsymbol{W}\boldsymbol{Y}_P - \underline{\boldsymbol{H}}^{\mathsf{T}}\right)\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{D}^{\frac{1}{2}} \right\|_{\mathsf{F}}^2 \tag{106}$$

$$\text{subject to: } (1)\ \boldsymbol{W} = \frac{g_l}{q^2}\underline{\boldsymbol{h}}_0\boldsymbol{y}_0^{\mathsf{H}} + g_c\boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y \tag{107}$$

$$(2)\ \boldsymbol{P}_C\boldsymbol{P}_C^{\mathsf{H}} = \boldsymbol{I}, \tag{108}$$

where $||\cdot||_{\mathsf{F}}$ denotes the Frobenius norm, $\boldsymbol{Y}_P = [\boldsymbol{y}_P(\Omega_1)\ \cdots\ \boldsymbol{y}_P(\Omega_J)]$, and $\boldsymbol{D} = \operatorname{diag} d(\Omega_j)$. The sampling positions on the sphere Ω_j and the quadrature weights $\boldsymbol{G} = \operatorname{diag} g(\Omega_j)$ have to be carefully chosen in order to yield a good approximation. Inserting (107) into the cost function and expanding the Frobenius norm, while constant terms can be dropped, yields

$$\min_{\boldsymbol{P}_C} \operatorname{tr}\left(g_c^2 \boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y\boldsymbol{M}\boldsymbol{N}_y^{\mathsf{H}}\boldsymbol{P}_C^{\mathsf{H}}\boldsymbol{K}\boldsymbol{K}_C\right) \tag{109}$$

$$+2\,\mathfrak{Re}\left\{\operatorname{tr}\left(\frac{g_l g_c}{q^2}\,\boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y\boldsymbol{M}\boldsymbol{y}_0\underline{\boldsymbol{h}}_0^{\mathsf{H}}\right)\right\} \tag{110}$$

$$-2\,\mathfrak{Re}\left\{\operatorname{tr}\left(g_c\boldsymbol{K}\boldsymbol{K}_C\boldsymbol{P}_C\boldsymbol{N}_y\boldsymbol{Y}_P\boldsymbol{G}\boldsymbol{D}\underline{\boldsymbol{H}}^*\right)\right\}, \tag{111}$$

where $\boldsymbol{M} = \boldsymbol{Y}_P\boldsymbol{G}\boldsymbol{D}\boldsymbol{Y}_P^{\mathsf{H}}$. In general, this can not be solved analytically, but if $\boldsymbol{D} \propto \boldsymbol{I}$, the first term is constant under the identity constraint since $\boldsymbol{N}_y\boldsymbol{M}\boldsymbol{N}_y^{\mathsf{H}} = \boldsymbol{I}$, the second term vanishes since $\boldsymbol{N}_y\boldsymbol{y}_0 = \boldsymbol{0}$. The problem is thus reduced to the residual problem

$$\max_{\boldsymbol{P}_C} \operatorname{tr}\left(\boldsymbol{P}_C\boldsymbol{A}\right) \tag{112}$$

$$\text{subject to } \boldsymbol{P}_C\boldsymbol{P}_C^{\mathsf{H}} = \boldsymbol{I}, \tag{113}$$

where

$$\boldsymbol{A} := \boldsymbol{K}\boldsymbol{K}_C\boldsymbol{N}_y\mathring{\underline{\boldsymbol{H}}}_P^*. \tag{114}$$

It follows that $\boldsymbol{P}_C$ is the unitary matrix that maximizes (112) is the conjugate transpose of the closest unitary rectangular matrix to $\boldsymbol{A}$ [18]. Via singular value decomposition of

$$\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{\Sigma}_A\boldsymbol{V}_A^{\mathsf{H}} \tag{115}$$

it can be found that

$$\boldsymbol{P}_C = \boldsymbol{U}_A\begin{bmatrix}\boldsymbol{v}_{A,1} & \boldsymbol{v}_{A,2}\end{bmatrix}. \tag{116}$$

4 Implementation

Real-time audio applications require that all computations performed on the input samples are finished within a certain interval of time. In audio applications, usually a callback function is used which receives a block of input samples as a parameter and returns the output samples. Failing to return the block of samples within the time limit causes output underflow or input overflow, which result in audible artifacts.

It is thus essential to avoid any operations that take an unpredictable amount of time during the execution of the callback function. Such operations may be, e.g., memory allocations and deallocations, or file read and write operations. Furthermore, resources required by the callback function must not be blocked by other threads.

In order to reduce the computation time as much as possible, the time-consuming optimization is performed off-line. The plug-in is provided with a functionality to pre-compute the weights for custom HRTFs and save them to a *preset file* which can be loaded later on. Furthermore, since each frequency bin undergoes the same computations, speed can be easily improved by parallelization. This can be achieved by splitting the range of frequency bins into a number of partitions which are processed each in a separate thread. An alternative way to achieve parallelization could be to make use of the SIMD functionality of modern processors, or to perform part of the computations on the GPU. However, the current implementation confines itself to the use of threads. The number of partitions that can be processed in parallel is limited by the number of CPU kernels. However, this maximum can not be guaranteed, mainly because the CPU may be busy processing other threads running at the same time, including the plug-in main thread, the host DAW thread, or threads of other applications and the operating system.

4.1 Pre-Computation of the Filter Weights

In Section 3, the optimal filter weights in a least squares sense were derived. The optimization problem needs to be solved for each frequency bin of each time frame. One way to reduce the amount of computations per time frame is to pre-compute the filter weights. Since the weights constitute a function on a sphere surface, it is convenient to express them in the spherical harmonic domain. Compared to saving the weights evaluated for a fixed coordinate grid directly, the SH representation brings the advantage of implicit interpolation when the SH coefficients are evaluated for a certain direction. Thus, they can be evaluated continuously, instead of having to map the DOA to the closest point of the grid or to employ an additional interpolation step.

Because it is sufficient to produce a perceptively sufficient result, perfect reconstruction is not necessarily required. The expansion order Q' of the weights may therefore be chosen lower than the expansion order Q of the HRTFs in many cases, which further reduces the amount of computations. To gain flexibility, the weights are stored at a sufficiently high order, while the user is free to choose a lower order, e.g., if a lower order already produces perceptively accurate results, or if the computational load is too high.

Choosing an order which is too low may lead to inaccurate results. We observed that if the order is reduced below 5-6, the decrease of spatial resolution is clearly perceptible. Very high orders, however, do not seem to result in significant improvements, especially compared against the increase of computational load.

The weights are computed for the full sphere at a dense sampling grid. The SH transform is well-conditioned if a sufficiently dense spherical t-design is used as a grid.

Recall that one time-frequency bin of the estimated binaural signal vector $\hat{\boldsymbol{x}}$ originating from a sound source located at the DOA Ω_0 is given by the product

$$\hat{\boldsymbol{x}} = \boldsymbol{W}_0 \, \boldsymbol{z} = \begin{bmatrix} \boldsymbol{w}_{\mathsf{L},0} \\ \boldsymbol{w}_{\mathsf{R},0} \end{bmatrix} \boldsymbol{z}, \tag{117}$$

where the $(2 \times N_P)$ matrix $\boldsymbol{W}_0$ contains the beamformer weights optimal for Ω_0, and $\boldsymbol{z}$ is the order-P Ambisonic input signal vector.

In order to find a function on the sphere that can be transformed into the SH domain, the optimization problem in Section 3 is solved for a set of L_{eval} points, yielding two $(L_{\text{eval}} \times N_P)$ matrices $\boldsymbol{W}_{\mathsf{L}}$ and $\boldsymbol{W}_{\mathsf{R}}$. The SH transforms of these matrices are given by

$$\mathring{\boldsymbol{W}}_{\mathsf{L}} = \boldsymbol{Y}_{Q'}^{\dagger} \boldsymbol{W}_{\mathsf{L}} \tag{118}$$

$$\mathring{\boldsymbol{W}}_{\mathsf{R}} = \boldsymbol{Y}_{Q'}^{\dagger} \boldsymbol{W}_{\mathsf{R}}, \tag{119}$$

and have the dimensions $(N_{Q'} \times N_P)$. They are given by the product of the partial matrices for the left and right channel and $\boldsymbol{Y}_{Q'}^{\dagger}$, which is the pseudoinverse of

$$\boldsymbol{Y}_{Q'}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{y}_{Q'}(\Omega_1) & \boldsymbol{y}_{Q'}(\Omega_2) & \cdots & \boldsymbol{y}_{Q'}(\Omega_{L_{\text{eval}}}) \end{bmatrix}. \tag{120}$$

The estimated binaural signal $\hat{\boldsymbol{x}} = \boldsymbol{W}(\Omega) \, \boldsymbol{z}$ can now be computed for arbitrary Ω, where

$$\boldsymbol{W}(\Omega) := \begin{bmatrix} \boldsymbol{w}_{\mathsf{L}}(\Omega) \\ \boldsymbol{w}_{\mathsf{R}}(\Omega) \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_{Q'}^{\mathsf{T}}(\Omega) \mathring{\boldsymbol{W}}_{\mathsf{L}} \\ \boldsymbol{y}_{Q'}^{\mathsf{T}}(\Omega) \mathring{\boldsymbol{W}}_{\mathsf{R}} \end{bmatrix} \tag{121}$$

is the SH domain weighting matrix evaluated at a single point Ω with dimensions $(2 \times N_P)$.

Direct evaluation of the spherical harmonic polynomials involves expensive trigonometric computations. It is more efficient to make use of recurrence relations and to evaluate a pre-factored form, optionally using SIMD instructions, as shown by the authors in [19]. The authors also provide a source code generator[1] which was used to generate the SH evaluation functions the plug-in uses.

[1] `http://jcgt.org/published/0002/02/06/`

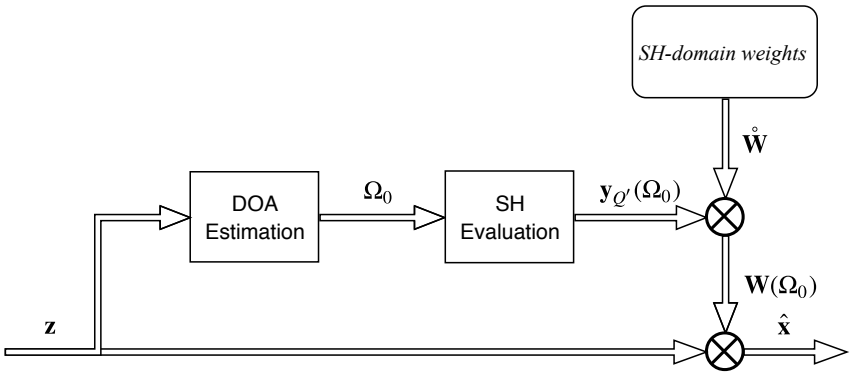


Figure 5: Block diagram of all computations performed at runtime.

4.1.1 Perceptual and Numeric Evaluation

Fig. 5 shows the signal flow at runtime for precomputed weights available in the SH domain. This is the core of the plug-in and the computations are performed on every STFT frame.

The DOA is estimated analogously to DirAC (see Sec. 2.4.1) based on the intensity vector. Informal listening showed that, in contrast to DirAC, no short-time averaging of the instantaneous intensity vector is required because no artifacts were produced.

The weighting coefficients are stored as time-domain impulse responses which are zero-padded according to the required FFT size and resampled according to the sampling frequency setting in the DAW when loaded. The expansion order can be freely chosen by the user. The effect of the SH transformation and successive dynamic evaluation at truncated order depending on the user setting of the accuracy of the filter coefficients is being investigated in the following. Fig. 6 compares the original impulse responses before SH transformation $\boldsymbol{w}_{\text{ref}}$ and the impulse responses $\boldsymbol{w}_{\text{eval}}$ after SH transformation and evaluation at the exact same coordinates the filter weights were optimized for. The weights were generated up to order $Q' = 20$ and for B-Format with HRTFs obtained from a spherical far-field HRIR of the Neumann KU 100 artificial head [11] with 2702 measurement positions (`HRIR_L2702.sofa`). The impulse responses are compared by correlation

$$\rho_{\boldsymbol{w}} = \frac{\boldsymbol{w}_{\text{eval}}^{\text{H}}\, \boldsymbol{w}_{\text{ref}}}{||\boldsymbol{w}_{\text{eval}}||\, ||\boldsymbol{w}_{\text{ref}}||} \tag{122}$$

in order to obtain a scalar similarity measure between 0 and 1 that can be displayed on a sphere surface. $\rho_{\boldsymbol{w}}$ can be interpreted as the angle between the impulse responses, where $\rho_{\boldsymbol{w}} = 1$ indicates that the impulse responses are parallel (high similarity) and $\rho_{\boldsymbol{w}} = 0$ indicates orthogonality (no similarity). The sphere surface is mapped onto a 2-dimensional plane using the azimuth as the horizontal and the elevation as the vertical axis, while the horizontal axis was contracted depending on the elevation. The data

points were linearly interpolated. The rows of Fig. 6 show the minimum values of the correlation of the left and right impulse responses for the four components of a first-order input signal and increasing expansion order up to $Q' = 20$. Processing with $Q' = 20$ is close to the upper limit for processing in real-time, but still feasible depending on the FFT size, while $Q' = 14$ is unproblematic. For the lower expansion orders deviations become visible. Values below 0.9 are not individually resolved because $\rho_{\boldsymbol{w}} \leq 0.9$ was considered insufficient. Since plots for $Q' = 14, 20$ show a high correlation, a sufficient resolution is feasible. Reducing the order can be interpreted as smoothing the filter weights. The choice of the expansion order is left to the user so that he or she can make a compromise between accuracy and required computational power if necessary.

4.2 Handling of Incomplete HRTF Data

HRTFs are usually measured for a grid of discrete positions around a subject's or artificial head. In order to interpolate the HRTF set for arbitrary positions, the spherical harmonic transform (SHT) can be employed. The reliability of the SHT depends on the number of grid points and their distribution on the sphere. Appropriate sampling can not be ensured for measured HRTFs in general, as, especially if a human subject is used, certain angles can not be covered (usually a cap at the bottom of the sphere) and the available amount of time for the measurement is limited. Since the underlying continuous data is unknown and not available for comparison, the quality of the approximation must be judged on the basis of similarity of the original data and the SHT evaluated for the same positions, and how plausible the values in between are in a perceptual sense.

The discrete spherical harmonic transform of an HRTF set is an approximation of the unknown continuous analysis integral (cf. (6))

$$h_{nm} = \int_{\Omega \in \mathcal{S}^2} h(\Omega) Y_n^m(\Omega) \mathrm{d}\Omega. \tag{123}$$

By the choice of an appropriate coordinate grid $\Omega_j \in \mathcal{M}$, and, if necessary, quadrature weights $\boldsymbol{G} = \mathrm{diag}\{g_j\}$ for each sample on the sphere, the order-N SHT of an HRTF set $\boldsymbol{H} = [\boldsymbol{h}(\Omega_j)]_{\Omega_j \in \mathcal{M}}$, where each $\boldsymbol{h}(\Omega_j)$ consists of the left and right HRTF coefficient for direction Ω_j, is given by

$$\mathring{\boldsymbol{H}}_N = \boldsymbol{Y}_N \boldsymbol{G} \boldsymbol{H}, \tag{124}$$

where $\boldsymbol{Y}_N = [\boldsymbol{y}_N(\Omega_j)]_{\Omega_j \in \mathcal{M}}$ contains vectors of spherical harmonic coefficients up to order N for all Ω_j. The pseudoinverse of $\boldsymbol{Y}_N$ may be used instead of determining the integration weights, yielding the least squares solution:

$$\mathring{\boldsymbol{H}}_N = \boldsymbol{Y}_N^\dagger \boldsymbol{H} \tag{125}$$

In case of inappropriate sampling, the pseudoinverse may be ill-conditioned, leading to numerical instability. Since the plug-in should not only be compatible with dense artificial

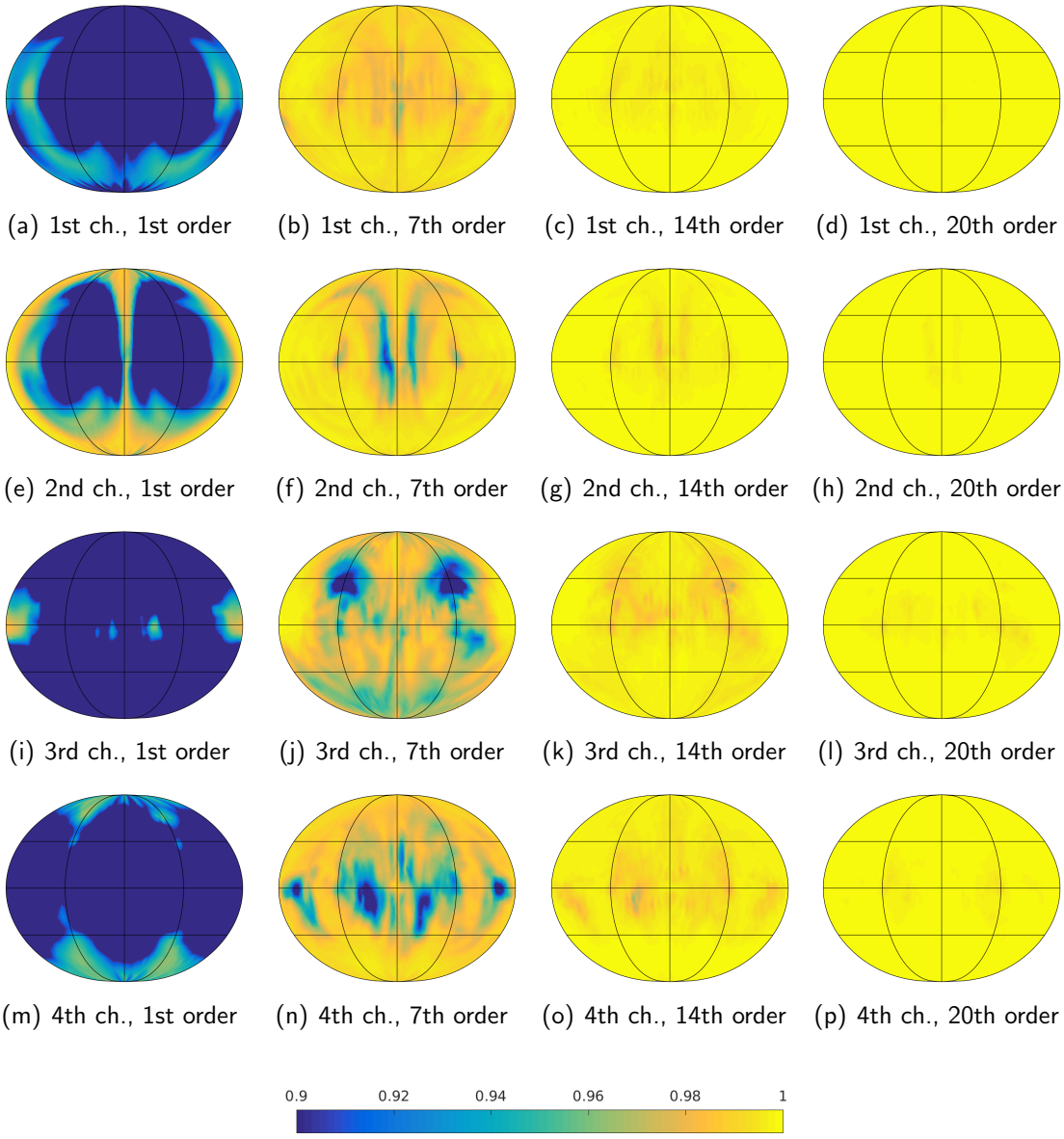
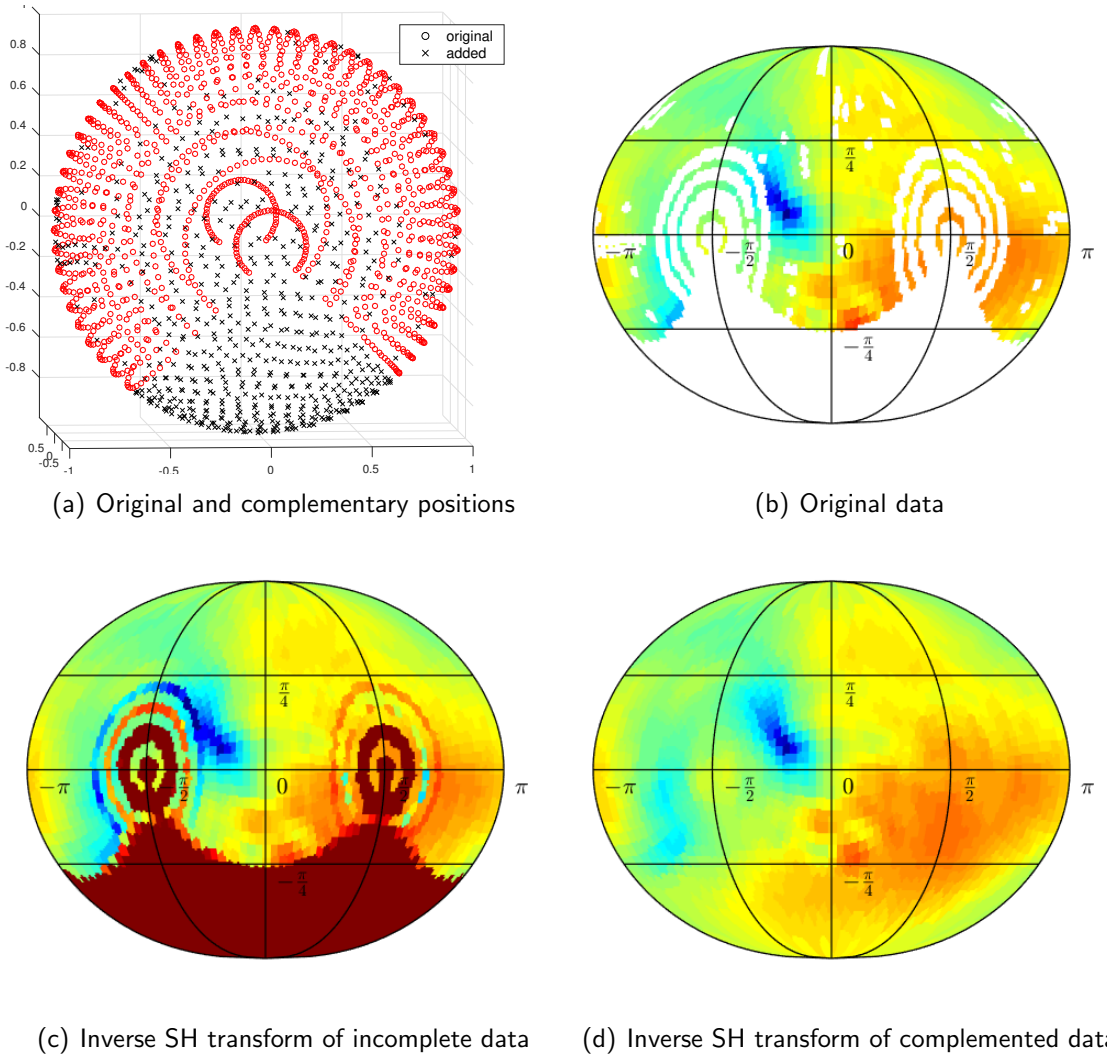


Figure 6: Correlation (angle) ρ_w between original and SH-evaluated impulse responses (minimum of left and right channel). Values below 0.9 are mapped to 0.9.

head HRTF sets, but also possibly sparse HRTF sets measured with a human-subject, a way must be found to cope with the difficulties arising from incomplete HRTF data.

One approach is to regularize the least squares fit in (125), as shown in [20]. The authors of [21] propose a very simple alternative to that approach. The original data is transformed into the SH domain using the least squares method with low order, e.g., $N = 3$. In this case, the problem should be well-behaved since pseudoinverse is expected to be well-conditioned. In a next step, new samples are generated by evaluating the obtained low-order SH coefficients for any desired grid points in the unknown regions. The incomplete HRTF set is complemented with these new samples and a higher-order

SHT can be performed on the combined set. Of course, this method can not reconstruct missing data, but it ensures numerical stability, has little effect on the measured samples, and does not lead to severe artifacts if the SHT is evaluated in the unknown regions.



(a) Original and complementary positions

(b) Original data

(c) Inverse SH transform of incomplete data

(d) Inverse SH transform of complemented data

Figure 7: Impact of complementary points on the SH transform of an HRTF (`subject_003.sofa`, right ear) for the $1\,\mathrm{kHz}$ band.

To determine the locations where new samples should be added, this implementation compares the original grid with a dense t-design. At each coordinate point of the t-design which is further away from any measurement position than a fixed maximum angle, a new sample is generated. The maximum distance the implementation uses was set to $3°$.

Fig. 7 shows an example. An HRTF from the CIPIC HRTF Database [22] (`subject_003.sofa`) with 1250 source positions was used. As it can be seen in Fig. 7(a), the original source positions (*red* $\circ$) are located on an arc that is rotated around the interaural axis. The measurement grid is less dense close to the rotation axis, and a large surface element at the bottom is not covered at all. Fig. 7(b) shows the magnitude of the right HRTF at $1\,\mathrm{kHz}$ in dB. The data points are interpolated using the nearest-neighbor method for smoother visualization. The blank areas correspond to the points that will be added in the interpolation step, i.e., all coordinates from a t-design ($t = 57$ and 1720 points) which are further away than 3° from any original data point, as shown in Fig. 7(a) (*black* $\times$). Fig. 7(c) shows the original HRTF data after SH transformation and successive evaluation with order 20. The SH coefficients were evaluated for the combined set of the original and the complementary points. It is clearly visible that the multiplication by the ill-conditioned matrix inverse in (125) leads to excessive magnitude levels at the missing areas. Obviously, this is not a desirable behavior. Fig. 7(d) shows, in contrast, that if the presented method is applied and new interpolated points are added, the magnitude level stays within the expected range with smooth transitions to the surrounding original values, and the original values are not noticeably altered.

4.3 File Format

The common exchange format for HRTF measurements is the SOFA format (Spatially Oriented Format for Acoustics). The plug-in should be compatible with files following the SOFA convention[2] [23]. The format is based on NetCDF which is an exchange format for scientific data that provides storage functionality for multi-dimensional data and corresponding metadata. In the case of SOFA files, the HRIRs are stored together with the respective source and receiver coordinates, and the definition of the coordinate system.

It is convenient to use NetCDF also for storing the precomputed weights. The expansion order is provided among the metadata, and the plug-in can adapt to files generated for arbitrary orders.

For reading and writing NetCDF files the `netcdf-cxx-4.2` library[3] was used.

4.4 Class Structure

The plug-in is written in C++ using the JUCE framework[4]. JUCE allows to build various plug-in formats as well as standalone applications for Windows, Mac OS X, and Linux platforms. It provides a ready-to-use environment for audio processing including the audio callback and a graphical user interface (GUI).

A minimal JUCE plug-in consists of the *PluginProcessor* (audio processing) and the

[2] `https://www.sofaconventions.org`

[3] `https://www.unidata.ucar.edu/downloads/netcdf/netcdf-cxx/index.jsp`

[4] `https://www.juce.com/`

PluginEditor (user interface). Several classes were added in the current implementation. The following is an overview of the classes. An excerpt of the inheritance diagram of the `BinauralAudioProcessor`, which is the implementation of the *PluginProcessor* of this plug-in, is provided in Fig. 8.

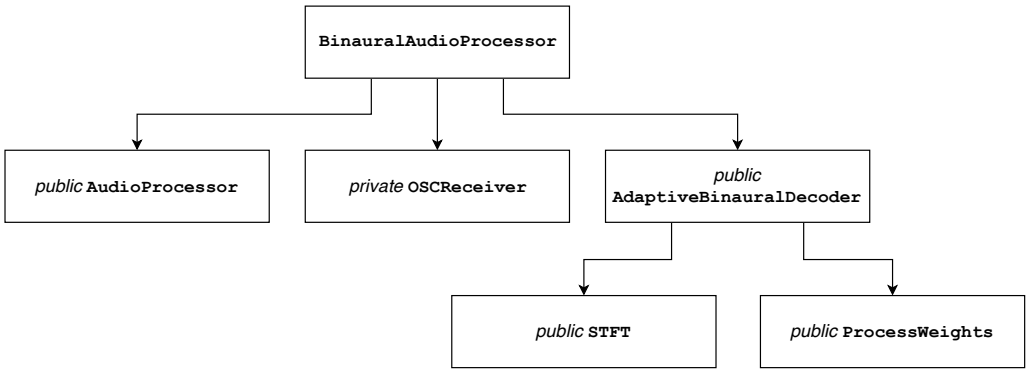


Figure 8: Inheritance diagram of the `BinauralAudioProcessor`.

BinauralAudioProcessor This is the main processing class. It inherits general audio processing functionality from `AudioProcessor`. This includes, e.g., the audio callback. It is a generic implementation that can be wrapped as different plug-in formats or standalone applications. The implementation of the binaural renderer is inherited from `AdaptiveBinauralDecoder`.

BinauralPluginEditor Implements the graphical user interface (GUI), inherited from the JUCE class `AudioProcessorEditor`.

AdaptiveBinauralDecoder This is the actual implementation of the Ambisonics-to-binaural rendering algorithm. It holds the algorithm parameters, and performs all steps of the binaural rendering algorithm. First, the input audio data are passed to an STFT (short-time Fourier transform) routine inherited from the `STFT` base class. Frequency-domain audio processing is implemented in a pure virtual method of `STFT` that has to be overridden by the inheriting class. In this method, the DOA is estimated and a vector of spherical harmonics is evaluated at the DOA for each frequency bin. Subsequently, the filter weights are evaluated at the DOA and multiplied with the input data. The corresponding methods are inherited from the second base class, `ProcessWeights`, which furthermore provides functionality for loading the filter weights.

STFT A framework for performing the short time Fourier transform using the FFTW library [24].

ProcessWeights Used to load and prepare the filter weights, and to multiply the filter weights evaluated for the DOA with the audio input data. Preparing the filter weights refers to loading all coefficients up to the specified SH order, resampling, zero padding, and FFT.

GenWeights This class is used for loading an HRTF set from a SOFA file, computing, and storing the filter weights. It inherits from `ThreadWithProgressWindow`. A `GenWeights` object launches a thread for parallel computation of the weights without halting the main processing thread (DAW). When the inherited `run` method is called, a window is opened with a progress bar which is continuously updated to give the user feedback about the state of the computations.

4.5 Usage Example

The following steps are an example of the workflow for using the plug-in. Note that, as the plug-in is continuously under development, the GUI and certain features may change in the future. The present GUI is depicted in Fig. 9.

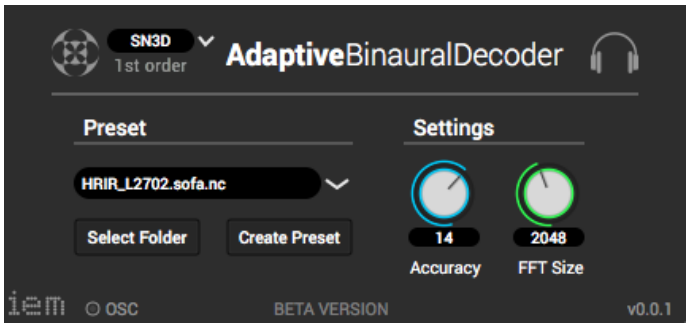


Figure 9: Graphical user interface of the plug-in.

1. Open a DAW of your choice and load the plug-in into a 4-channel track or bus.
2. Choose appropriate audio material for processing, i.e., first-order Ambisonics (B-Format). The plug-in works with ACN channel ordering and SN3D or N3D normalization. If the channel ordering and normalization differ from this convention, you may use a converter plug-in.
3. The plug-in requires a preset file which contains the filter weights generated for a HRTF set.
4. To create a preset file, click the *Create Preset* button and open a HRTF set. Valid HRTF sets follow the SOFA convention and have the file ending `.sofa`.
5. To load a preset file, navigate to the directory where your preset files are stored by clicking the *Select Folder* button. Choose a preset file using the combo box which lists all preset files with the ending `.nc` in the current directory.

6. Adjust the plug-in settings if necessary. The optimal *FFT Size* parameter may differ for different types of audio material, e.g., music or speech. The *Accuracy* parameter may be chosen between 1 and the maximum evaluation order used for generation of the current preset. A lower value than the maximum may be sufficient depending on the audio material and will lead to less CPU usage. However, note, that choosing a too low value will lead to inaccurate reproduction.

7. *Hint:* You can read the data and metadata of SOFA and preset files using software tools like, e.g., `ncdump`.

5 Conclusion

Binaural rendering of Ambisonic signals can be interpreted as the approximation of the HRTF directional patterns by means of a set of basis functions (spherical harmonics). The spatial complexity of HRTFs increases with frequency. Consequently, a significant amount of the total energy is contained in higher order modes. Modeling the directivity patterns of the ears (HRTFs) requires spherical harmonic orders up to 35. The unconstrained least squares binaural decoder introduces sound colorations due to the modal order mismatch between a lower-order Ambisonic input signal and the HRTF. Moreover, lower-order Ambisonics yields poor directional resolution.

Different strategies exist to cope with these problems. Sound colorations can be reduced by aliasing higher-order modes into lower orders by spatial sub-sampling [7], or by applying global diffuse field equalization [8]. The spatial complexity of HRTFs at high frequencies can be reduced by removing the linear phase [10] above the frequency above which phase information is not evaluated for localization anymore. The magnitude least squares method [9] tries to find the optimal phase modification while minimizing the magnitude squared error between the decoder and the HRTFs. Note that the listed methods operate independently of the input signal.

Various methods try to obtain a signal with enhanced resolution from B-Format input. Based on the assumption of a single direct source per frequency bin plus a diffuse component, Directional Audio Coding (DirAC) [12] [13] separates the input signal into a direct and a diffuse stream, which are weighted according to the estimated diffuseness parameter. The direct component is encoded as point sources in the direction of the estimated direction of arrival (DOA), while the diffuse component is a decorrelated version of the input. Critical parameters are the estimated DOA and diffuseness, and the time constants for averaging. High Angular Resolution Plane Wave Expansion (HARPEX) [14] decomposes the input into four plane waves per frequency bin which are then encoded with higher order. The authors of [16] use compressed sensing techniques to yield a plane wave decomposition with a low number of discrete sources that can be encoded in higher order by convolution with a time-domain filter matrix.

The heart of this thesis was the presentation of a recently proposed parametric binaural rendering method for Ambisonics signals. This method uses signal-dependent linear filtering in the frequency domain. Similar to DirAC, it is assumed that a single direction of arrival (DOA) per frequency bin and time frame substantially contributes to localization. The sound field is modeled as a single plane wave from the direction of arrival (DOA) plus diffuse sound. According to this model, a complete description of the sound field is given by the DOA and complex amplitude of each of the point sources, and the inter-aural covariance matrix of the diffuse component.

The proposed method is a signal-dependent extension to the least-squares decoder. Based on the assumption of a single predominant direction of arrival (DOA) per time/frequency bin, a constrained beamformer is directed towards the DOA. The beamformer weights, which are the weights of a linear combination of spherical harmonic functions of the same order as the input signal, are subject to a doubly

constrained optimization problem. While the error between the directional response of the beamformer and the HRTF coefficients is minimized in the least-squares sense for all directions on the sphere, the constraints must be exactly met. The constraints demand that (i) the transfer function of the beamformer in direction of the most dominant plane wave must be identical to the HRTF for the respective direction, and (ii) the covariance matrix of the estimated binaural signal is identical to the covariance matrix of the desired signal. In contrast to DirAC, there is no need to estimate the signal-to-diffuse ratio, or to generate a decorrelated diffuse signal because these parameters are implicitly captured by the covariance constraint.

In the course of this thesis, a plug-in was developed that implements the new method. The plug-in has two main functionalities. The first and most important task is, of course, binaural rendering of Ambisonics input signals in real-time. Although the method works for arbitrary input orders, we decided to limit the input to first-order Ambisonics since the computational load massively increases for higher orders, while the advantage over static binaural rendering methods, like, e.g., [9], decreases. The second task is the pre-computation of the filter weights, as the optimization process would take too long to be performed in real-time. The weights are only signal-dependent with regard to the DOA. Hence, the weights are precomputed for a closely spaced grid of azimuth and elevation angles on the sphere surface and transformed to the spherical harmonic domain subsequently. If the weights are computed for a sufficiently dense grid, the transform is usually well-conditioned. For each direction in this grid, a binaural HRTF must be known. However, this can not be ensured, especially for HRTFs resulting from measurements with human subjects. Therefore, the plug-in compares the HRTF grid to a dense t-design and adds new interpolated grid points in insufficiently covered regions (or maybe rather *extrapolates* to the unknown regions) using the method presented in [21]. The HRTF set can now be represented in the SH domain at a sufficiently high order and evaluated for the desired directions.

While the described rendering method appears to yield excellent localization and very little artifacts, future work should concern subjective evaluation in a listening experiment. The errors made in DOA estimation and interpolation of the SH transformed filter weights need to be characterized by an objective measure. Furthermore, reliability of the computed weights depending on the number and distribution of HRTF measurement points could be investigated. An interesting research question would be whether the new method can be applied for parametric Ambisonics decoding for microphone arrays. A possible feature to complement the plug-in could be to use measured directivity patterns of the microphones used for recording the input signals, instead of implicitly assuming ideal microphone characteristics. This could include estimation of the beamformer weights for custom microphone setups. In addition, headphone equalization could be implemented in order to further improve localization by reducing the influence of the transfer path of the headphones.

References

[1] P. M. Giller and C. Schörkhuber, "A super-resolution ambisonics-to-binaural rendering plug-in," in *Conf.: Fortschritte der Akustik, Rostock*. DAGA, 2019.

[2] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of ambisonic recordings," in *Proc. of the 2nd Int. Conf. on Spatial Audio, Erlangen*, 2014.

[3] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[4] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. on speech and audio process.*, vol. 13, no. 1, pp. 135–143, 2005.

[5] C. An, X. Chen, I. Sloan, and R. Womersley, "Well conditioned spherical designs for integration and interpolation on the two-sphere," *J. Numerical Analysis*, vol. 48, pp. 2135–2157, 2010.

[6] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2219–2236, 2002.

[7] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[8] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. of the Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096, 2017.

[9] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Conf.: Fortschritte der Akustik, Munich*. DAGA, 03 2018.

[10] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," in *J. Acoust. Soc. Am.*, 2018.

[11] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proc. of the DAGA*, 2013, p. 29.

[12] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[13] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Appl. of Signal Process. to Audio and Acoust.* IEEE, 2009.

[14] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. of the 2nd Int. Symp. on Ambisonics and Spherical Acoust.*, 2010, pp. 6–7.

[15] N. Barrett and S. Berge, "A new method for B-Format to binaural transcoding," in *Audio Eng. Soc. Int. Conf.: Spatial Audio: Sense the Sound of Space*. Audio Eng. Soc., 2010.

[16] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.* IEEE, 2012, pp. 385–388.

[17] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Inter. Conf. on Immersive and Interactive Audio*. Audio Eng. Soc., 2019.

[18] C. Schörkhuber and R. Höldrich, "Ambisonic microphone encoding with covariance constraint," in *Int. Conf. on Spatial Audio*, 2017.

[19] P.-P. Sloan, "Efficient spherical harmonic evaluation," *J. of Comput. Graph. Techn. (JCGT)*, vol. 2, no. 2, pp. 84–83, September 2013.

[20] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. IEEE, 2009, pp. 257–260.

[21] J. Ahrens, M. R. Thomas, and I. Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Signal & Information Process. Assoc. Annu. Summit and Conf.* IEEE, 2012.

[22] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Workshop on the Appl. of Signal Process. to Audio and Acoust.* IEEE, 2001, pp. 99–102.

[23] P. Majdak *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Proc. of the 134th Conv. of the Audio Eng. Soc.* Audio Eng. Soc., 2013.

[24] M. Frigo and S. G. Johnson, "FFTW: an adaptive software architecture for the FFT," in *Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Process.*, vol. 3, May 1998, pp. 1381–1384 vol.3.